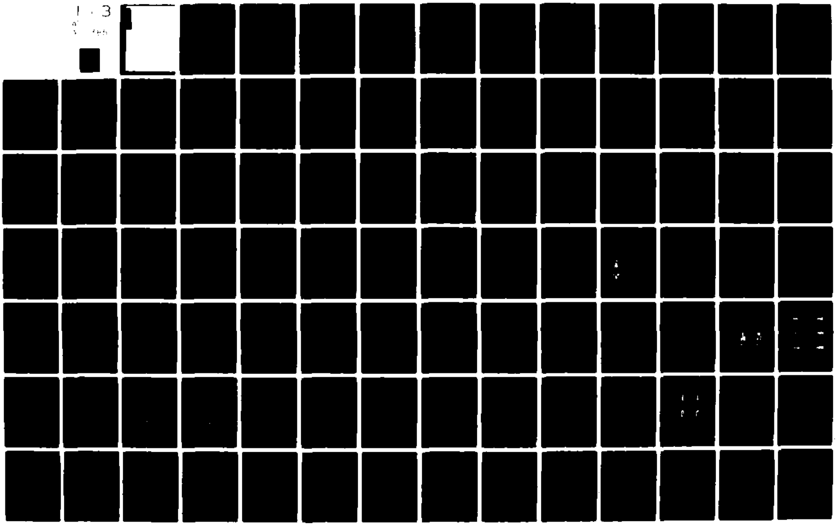
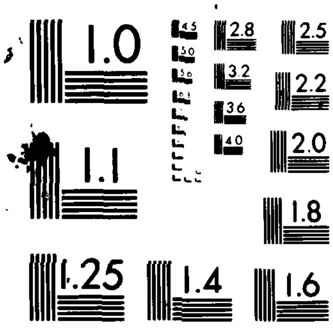
AD-A111 385   HASKINS LABS INC NEW HAVEN CT   F/8 6/16
STATUS REPORT ON SPEECH RESEARCH, JULY-DECEMBER 1981.(U)
1981   PHS-HD-01994
UNCLASSIFIED   SR-67/68(1981)   NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

SR-67/68 (1981)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 July - 31 December 1981

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development
Grant HD-01994
Grant HD-05677

National Institute of Child Health and Human Development
Contract N01-HD-1-2420

National Institutes of Health
Biomedical Research Support Grant RR-05596

National Science Foundation
Grant MCS79-16177
Grant PRF8006144
Grant BNS-8111470

National Institute of Neurological and Communicative Disorders and Stroke
Grant NS13870
Grant NS13617

National Institute of Education
Grant G-80-0178

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper,* Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Peter Alfonso*
Cinzia Avesani[2]
Thomas Baer
Fredericka Bell-Berti*
Catherine Best*
Gloria J. Borden*
Susan Brady*
Giuseppe Cossu[3]
Robert Crowder*
Carol A. Fowler*
Louis Goldstein*
Vicki Hanson
Katherine S. Harris*
Alice Healy*
Kiyoshi Honda[1]
Leonard Katz*
J. A. Scott Kelso
Andrea G. Levitt*
Isabelle Y. Liberman*
Leigh Lisker*
Virginia Mann*
Charles Marshall
Ignatius G. Mattingly*
Nancy S. McGarr*
Lawrence J. Raphael*
Bruno H. Repp
Philip E. Rubin
Elliot Saltzman
Donald P. Shankweiler*
Michael Studdert-Kennedy*
Betty Tuller*
Michael T. Turvey*
Mario Vayra[2]
Robert Verbrugge*

Technical and Support Staff

Eric L. Andreasson
Margo Carter
Elizabeth P. Clark
Vincent Gulisano
Donald Hailey
Terry Halwes
Sabina D. Koroluk
Bruce Martin
Agnes M. McKeon
Nancy O'Brien
Marilyn K. Parnell
Susan Ross*
William P. Scully
Richard S. Sharkany
Leonard Szubowicz
Edward R. Wiley
David Zeichner

Students*

Suzanne Boyce
Tova Clayman
Steven Eady
Jo Estill
Laurie B. Feldman
Carole E. Gelfer
Janette Henderson
Charles Hoequist
Robert Katz
Peter Kugler
Gerald Lame
Anthony Levas
Harriet Magen
Sharon Manuel
Suzi Pollock
Brad Rakerd
Daniel Recasens
Rosemarie Rotunno
Hyla Rubin
Judith Rubin
Arnold Shapiro
Suzanne Smith
Rosemary Szczesiul
Douglas Whalen
Deborah Wilkenfeld
David Williams

*Part-time
[1] Visiting from University of Tokyo, Japan
[2] Visiting from Scuola Normale Superiore, Pisa, Italy
[3] Visiting from Istituto Di Neuropsichiatria Infantile, Sassiri, Italy

CONTENTS

I. Manuscripts and Extended Reports

I. MANUSCRIPTS AND EXTENDED REPORTS

PHONETIC TRADING RELATIONS AND CONTEXT EFFECTS: NEW EXPERIMENTAL EVIDENCE FOR A SPEECH MODE OF PERCEPTION*

Bruno H. Repp

Abstract. This article reviews a variety of experimental findings, most of them obtained in the last few years, that show that the perception of phonetic distinctions relies on a multiplicity of acoustic cues and is sensitive to the surrounding context in very specific ways. Nearly all of these effects have correspondences in speech production, and they are readily explained by the assumption that listeners make continuous use of their tacit knowledge of speech patterns. A general auditory theory that does not make reference to the specific origin and function of speech can, at best, handle only a small portion of the wealth of phenomena reviewed here. Special emphasis is placed on several recent studies that obtained different patterns of results depending on whether identical stimuli were perceived as speech or as nonspeech. These findings provide strong empirical evidence for the existence of a special speech mode of perception.

INTRODUCTION

Speech is a specifically human capacity. Just as humans are uniquely enabled to produce the complex stream of sound called speech, one might suppose that they make use of special perceptual mechanisms to decode this complex signal. Of course, since speech is remarkably different from all other environmental sounds, it is highly likely that there are perceptual and cognitive processes that occur only when speech is the input. Otherwise, speech simply would not be perceived as what it is. To make sense, the question of whether speech perception is different from other forms of perception is best restricted to those aspects of speech that are not obviously unique, e.g., to its being an acoustic signal that can be described in the same physical terms as other environmental sounds. Then the question may be raised whether the perceptual translation of this acoustic signal into the sequence of discrete linguistic units that we experience (i.e., phonetic perception) requires the assumption of special mechanisms, or whether it can be reduced to a combination of auditory processes known to be involved also in

*A revised version is to appear in Psychological Bulletin.

Acknowledgment. Preparation of this paper was supported by NICHD Grant HD01994 and BRS Grant RR05596 to Haskins Laboratories. Valuable comments on an earlier draft were obtained from Carol Fowler, Alvin Liberman, Michael Studdert-Kennedy, Janet Werker, and an anonymous reviewer. My intellectual debt to Alvin Liberman must be evident throughout the paper.

[HASKINS LABORATORIES: Status Report on Speech Research SR-67/68 (1981)]

the perception and interpretation of nonspeech sounds. Even this modest question, however, presupposes that the linguistic categories applied by a listener, even though they are appropriate only for speech, are not unique in any essential sense but rather can be viewed as labels applied to specific auditory patterns. This assumption is probably wrong, but it must be granted now for the argument to proceed.

The precise nature of the processes and mechanisms that support phonetic perception has been the subject of much discussion. A number of speech researchers hold the view that speech perception is special in the sense that it takes account of the origin of the signal in the action of a speaker's articulatory system. This general view underlies the well-known motor theory of speech perception (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) as well as the theory of analysis-by-synthesis (Halle & Stevens, 1959). More recently, it has been fertilized and augmented by ideas derived from Gibson's (1966) theory of event perception (see, e.g., Bailey & Summerfield, 1980; Neisser, 1976; Summerfield, 1979), which postulates that all perception is directed towards the source of stimulation. While, in the Gibsonian framework, speech perception is not seen as basically different from the perception of other auditory (and visual) events, the special nature of the source (the human vocal tract) is acknowledged and emphasized. In this view, speech perception is special because the source of speech is special. There are other researchers, however, who would concur only with the second half of that statement (the special nature of the source), not with the first. They pursue the hypothesis that the processes involved in speech perception are essentially the same as those that support the auditory perception of nonspeech sounds, and that they operate without implicit reference to the sound-producing mechanisms that generate the speech signal. In this view, the specific complexity of speech perception results merely from the diversity and the number of elementary auditory processes required to deal with an intricately structured signal (see, e.g., Divenyi, 1979; Kuhl & Miller, 1978; Pastore, 1981; Schouten, 1980; Stevens, 1975). These two views are perhaps most clearly distinguished by their different orientations to the evolution of speech perception: Whereas, according to the first view, special perceptual processes evolved hand in hand with articulatory capabilities to handle the complex output of a speaker's vocal tract, the second view assumes that the vocal productions of early hominids were fitted into a mold created by the pre-existing sensitivities and limitations of their auditory systems.

Which of these two views is correct is, in part, an empirical question that rests on many possible sources of evidence, including the reactions to speech of animal and human infant subjects, traditional laboratory experiments, electrophysiological and clinical observations. In this review, I will focus on a set of recent attempts to demonstrate the peculiarities of speech perception in the laboratory, using normal adult human subjects. This kind of evidence has been, and continues to be, central to the argument, as it is easier to obtain, permits a variety of approaches, and is perhaps more readily interpreted than some of the other research. This is not to deny that some of the most crucial results will come from infant and animal experiments; however, this research characteristically lags one step behind the standard laboratory findings, and studies that extend the latest findings on college students' perception to other subject populations are just getting under way as this review is being written.

Less than a decade ago, a rich set of experimental data apparently supported the existence of a special speech mode of perception, distinct from other kinds of auditory perception. However, within a few years that support seems to have all but evaporated. The history of these events will be summarized and commented upon in the first part of the present paper. Since the main purpose of that section is to set the stage for the following review, my treatment of what are complex and often controversial issues will necessarily be somewhat sketchy and betray my biases. In the second part, new evidence--much of it collected over the last few years--will be reviewed and discussed. I will conclude that we have, once again, strong experimental support for a special phonetic mode of perception.

THE OLD EVIDENCE

In a well-known paper, Wood (1975) listed six laboratory phenomena that, at that time, seemed to provide strong converging evidence for the existence of special processes in speech perception. One phenomenon is the "phoneme boundary effect," which is commonly subsumed under the more general term, categorical perception. It is the finding that two speech stimuli are easier to discriminate when they can be assigned to different linguistic categories than when, though separatea by an equivalent physical difference, they are perceived as belonging to the same category. A second phenomenon is selective adaptation, the shift of the category boundary on a synthetic speech continuum following repeated presentation of one endpoint stimulus. Three other phenomena have to do with hemispheric specialization: the dichotic right-ear advantage, the right-ear advantage in temporal-order judgments of speech stimuli, and differences in evoked potentials from the two hemispheres in response to speech stimuli. A sixth phenomenon concerned asymmetric interference between auditory and phonetic stimulus dimensions in a speeded classification task. Many of the findings that Wood referred to under these headings have been excellently reviewed by Studdert-Kennedy (1976).

At the time the Wood and Studdert-Kennedy papers were written, all of the above-named phenomena seemed to be specific to speech; that is, they were apparently not obtained with nonspeech stimuli. However, a few years later, the picture had changed considerably. Using Wood's enumeration of findings as their starting point, both Cutting (1978) and Schouten (1980) reviewed more recent research using the various paradigms and concluded independently that there was no evidence for a special phonetic mode of perception. After that statement, the views of these two authors diverge: Cutting, a vigorous proponent of the Gibsonian view, argues for considering speech perception as merely one instance of auditory event perception (i.e., the perception of auditory events other than speech may be as--or nearly as--complex and special as speech perception), while Schouten, who represents a more narrowly psychophysical orientation, states rather bluntly that "speech and non-speech auditory stimuli are probably perceived in the same way" (p. 71), implying that all auditory perception rests on the same elementary processes.

The conclusions of both authors reflect their disillusion over the failure of a number of experimental techniques to produce results specific to speech. Since the relevant evidence has been competently reviewed by them and by others, I will deal with it only briefly, focusing primarily on its interpretation.

Categorical Perception

The "phoneme boundary effect" singled out by Wood (1975)--the enhanced discriminability across the phonetic boundary on a synthetic speech continuum--is merely one aspect of the complex phenomenon termed categorical perception. Other aspects are reduced context sensitivity in stimulus categorization and predictability of discrimination performance from identification scores (Repp, Healy, & Crowder, 1979). However, these latter two aspects have not been claimed to be specific to speech.

The speech-specificity of the phoneme boundary effect has been challenged on the grounds that analogous effects have been demonstrated for a variety of nonspeech continua: noise-buzz sequences (Miller, Wier, Pastore, Kelly, & Dooling, 1976), tone-onset-time (Pisoni, 1977), tone amplitude in the presence of a reference signal (Pastore, Ahroon, Baffuto, Friedman, Puleo, & Fink, 1977), visual flicker (Pastore et al., 1977), musical intervals (Burns & Ward, 1978), and amplitude rise-time (Cutting & Rosner, 1974). The results for the rise-time ("pluck"-"bow") continuum, which have been widely cited and followed up, and on which Cutting (1978) rested his whole argument, have recently been claimed to be artifacts due to faulty stimulus construction (Rosen & Howell, 1981), but the other findings appear to be solid. However, some of them are not very surprising. If a psychophysical continuum is chosen on which some kind of threshold is known to exist--such as the critical flicker fusion threshold--it is obvious that two stimuli from opposite sides of the threshold will be more discriminable than two stimuli from the same side. However, it does not follow that, therefore, the phoneme boundary effect on a speech continuum is also caused by a psychophysical boundary that happens to coincide with the phoneme boundary. The problem is that, in most cases, we have no good idea of what the psychophysical boundary ought to be. Moreover, a phoneme boundary effect may be caused by the phoneme boundary itself, as argued below. There are several reasons why the nonspeech studies referred to above have done relatively little to clarify the issue.

First of all, only results obtained with nonspeech stimuli that have something in common with speech are directly relevant to the question of whether a specific phoneme boundary falls on top of a psychoacoustic threshold. For example, the observations on the flicker fusion threshold (Pastore et al., 1977) cannot have any direct implications for speech perception. They show that categorical perception can occur in the nonspeech domain, but they do not prove that the causes are the same as in a particular speech case. Second, just how much certain nonspeech stimuli have in common with speech stimuli they are intended to emulate is a matter of debate. It is doubtful, for example, whether the relative onset time of two sinusoids (Pisoni, 1977) successfully simulates the distinction between a voiced and a voiceless stop consonant (cf. Pastore, Harris, & Kaplan, 1981; Pisoni, 1980; Summerfield, in press), or whether amplitude rise-time has much to do with the fricative-affricate distinction (Remez, Cutting, & Studdert-Kennedy, 1980). Third, even those nonspeech continua (such as noise-buzz sequences) that appear to copy a speech cue more or less faithfully yield results that, on closer insp^ction, are not in agreement with speech results. For example, individual listeners in the Miller et al. (1976) study showed boundaries as short as 4 msec on a noise-buzz continuum, which is much shorter than any boundaries for English-speaking listeners on the supposedly analogous voice-onset-time dimension (see, e.g., Zlatin, 1974). Note also that auditory thresholds may shift with extended practice in the laboratory, while linguistic boundaries ordinarily do not; this creates a problem for comparing the locations of the two. Fourth, and most significantly, the various comparisons of categorical perception of speech and supposedly analogous nonspeech stimuli generally have not taken into account the fact that there are multiple cues for each phonetic contrast and that perception of one cue, as it were, is not independent of the settings of other relevant cues. This issue, which has received particular attention only in the last few years, will be central to the second part of the present paper. Fifth, there are a variety of other factors that influence the locations of phonetic boundaries: language experience, speaking rate, stress, phonetic context, semantic factors, and so on. It remains to be shown that psychophysical thresholds are sensitive to all, or even some, of these variables (or their psychoacoustic analogs). Finally, we note that there are examples of category boundary effects on nonspeech continua that have no obvious psychophysical boundaries, viz., for musical intervals (Burns & Ward, 1978; Siegel & Siegel, 1977) or chords (Blechner, 1977; Zatorre & Halpern, 1979), which suggests that well-established categories of non-psychophysical origin may dominate perception.

In view of these arguments, one plausible account of the phoneme boundary effect remains that it arises from the use of category labels in discrimination. The best support for this hypothesis comes from studies that show a change in speech sound discriminability consequent upon a redefinition of linguistic categories for the same stimuli and the same listeners (e.g., Carden, Levitt, Jusczyk, & Walley, 1981). However, the use of category labels in discrimination is not unique to speech. The difference between speech and nonspeech in the discrimination paradigm probably rests on the nature of the categories: Phonetic categories are not only more deeply engrained than other categories, but they also bear a *special relation to the acoustic signal*. As Studdert-Kennedy (1976) has put it, speech sounds "name themselves." Therefore, linguistic categories will dominate perception in a discrimination task to a larger extent than nonspeech categories that frequently do not even exist pre-experimentally and, in those cases, merely serve to bisect the stimulus range. In addition, the acoustic distinctions underlying a category contrast may be finer in the case of speech and also are habitually ignored by listeners in a natural situation; therefore, they are more difficult to access in the context of a discrimination task.

The strongest evidence for the alternative hypothesis, that categorical perception of speech rests on nonlinguistic auditory discontinuities in perception, comes from research on human infants (for recent summaries, see Jusczyk, 1981; Morse, 1979; Walley, Pisoni, & Aslin, 1981) and nonhuman animals, particularly chinchillas (Kuhl, 1981; Kuhl & Miller, 1978). Allowing for the inevitable methodological differences and limitations, infants and (so far) chinchillas appear to perceive synthetic speech stimuli essentially the same way adults do, including superior discrimination of stimuli from different (adult) categories than of stimuli from the same category. These effects obviously reflect some "natural" boundaries, but it is not entirely clear whether these boundaries are strictly psychoacoustic in nature or whether they perhaps reflect some innate or acquired sensitivity to articulatory patterns. Even if they were psychoacoustic (this being the received interpretation of the infant and chinchilla findings), it is not certain that linguistic categories in fact depend on them. (See, however, Aslin & Pisoni, 1980, for a different view.) For example, children in the early stages of language use often are not able to make the perceptual distinctions infants seem to be capable of (Barton, 1980). There are still many open questions here. A fair assessment of the situation may be that the evidence on phoneme boundary effects neither strongly supports nor disconfirms the existence of a special speech mode of perception.

Selective Adaptation

The shifting of phoneme boundaries on a continuum by repeated presentation of stimuli from one category has been a favorite pastime of some speech perception researchers ever since Eimas & Corbit (1973) discovered the technique. (See Diehl, 1981, for a recent critical review.) In hindsight, this effort seems not to have been worthwhile. Since various kinds of nonspeech dimensions show selective-adaptation effects, it was to be expected that auditory dimensions of speech can be adapted as well. On the whole, this is what a score of studies show. The technique was considered interesting because it was thought to reveal the existence of "phonetic feature detectors" (Eimas & Corbit, 1973). However, the evidence for specifically phonetic effects in selective adaptation is scant, and what there is can probably be explained as shifts in response criteria or as effects of remote auditory similarity. Recent experiments by Sawusch and Jusczyk (1981) and particularly by Roberts and Summerfield (1981) strongly suggest that there is no phonetic component in selective adaptation at all, and that the effect takes place exclusively at a relatively early stage in auditory processing.

The concept of phonetic feature detectors is useless not only for the explanation of selective adaptation results (cf. Remez, 1979) but also from a wider theoretical perspective. None expresses this better than Studdert-Kennedy (in press) when he says that "we are dealing with tautology, not explanation. ... The error lies in offering to explain phonetic capacity by making a substantive physiological mechanism out of a descriptive property of language" (p. 225). For, "... the perceived feature is an attribute, not a constituent, of the percept, and we are absolved from positing specialized mechanisms for its extraction" (p. 227). Arguments such as these apply not only to the concept of phonetic feature detectors but to the concept of the feature detector in general. For these reasons, selective adaptation results cannot have any implications for or against the existence of a special speech mode.

Hemispheric Specialization

The empirical results supporting a hemispheric asymmetry for speech and language are rich and complex. While left-hemisphere advantages have been reported for certain kinds of nonspeech sounds, the evidence that speech processes are lateralized to the left hemisphere in the large majority of individuals is unassailable. It has been claimed, however, that precisely because certain nonspeech stimuli show similar effects, the lateralization of speech should be explained by a more general principle, e.g., by a specialization of the left hemisphere for auditory properties characteristic of speech (Cutting, 1978; Schouten, 1980), or by an analytic-holistic distinction between the two hemispheres (e.g., Bradshaw & Nettleton, 1981). In commenting on the last-named paper, Studdert-Kennedy (1981) has argued that the analytic-holistic hypothesis, while descriptively adequate, is ill-conceived from a phylogenetic viewpoint. Rather, since lateralization presumably evolved to support some behavior important to the species, it seems more likely that lateralization of motor control preceded or caused lateralization of speech processes, which in turn may be responsible for the superior analytic capabilities of the left hemisphere. The apparent specialization of the left hemisphere for certain auditory characteristics of speech may just as well be the consequence as the cause of the lateralization of linguistic functions. Thus, the existing evidence on hemispheric specialization can be interpreted in an alternative way that is more compatible with a biological viewpoint and that recognizes the special status of speech.

Other Laboratory Phenomena

Various other findings have been cited as evidence for or against a speech mode of perception. Thus, Wood (1975) mentions the phenomenon of asymmetric interference between auditory and linguistic dimensions in a speeded classification task. While this finding (whose methodological details need not concern us here) may reveal something about the auditory processing of speech, its implications for the existence of a special speech mode of perception are limited. Similar patterns of results have been obtained with nonspeech auditory stimuli (Blechner, Day, & Cutting, 1976; Pastore, Ahroon, Puleo, Crimmins, Golowner, & Berger, 1976), suggesting that the asymmetry has a nonphonetic basis.

Schouten (1980) adds to Wood's list two findings that seem to have even less bearing on the question of a phonetic mode of perception: A difference in the stimulus duration needed for correct order judgments with sequences of speech or nonspeech sounds (Warren, Obusek, Farmer, & Warren, 1969), and an asymmetry in the perception of truncated CV and VC syllables (Pols & Schouten, 1978). The first finding probably reflects the fact that speech stimuli are more readily categorized than nonspeech stimuli, while the second finding seems altogether irrelevant, having most likely a psychoacoustic explanation. It is a mistake to believe (as Schouten apparently does) that the "case against a speech mode of perception" is strengthened by various findings of auditory (nonphonetic) effects in speech perception experiments. Such effects are likely to occur for, after all, speech enters through the ears. The thesis of the present paper is, however, that these effects are relatively inconsequential for the linguistic processing of speech.

By focusing primarily on the experimental paradigms listed in Wood's (1975) article, Cutting (1978) and Schouten (1980) neglected a variety of other observations that suggest the existence of a speech mode of perception. Liberman et al. (1967) reviewed many properties that are peculiar to speech and seem to require special perceptual skills. Foremost among these properties is the invariance of phonetic perception over substantial changes in the acoustic information; consider the well-known /di/-/du/ example, which shows that the /d/ percept can be cued by radically different transitions of the second formant. To achieve the same classification without reference to the articulatory gesture common to /di/ and /du/, an exceedingly complex "auditory decoder" would be required.

Liberman et al. (1967) also noted that the formant transitions distinguishing /di/ and /du/ sound quite different from each other when they are presented in isolation and do not engage the speech mode. In fact, when second- or third-formant transitions are removed from a synthetic syllable and presented to one ear while the rest of the speech pattern is presented to the other ear, the transitions are found to do double duty: They are perceived as whistles or chirps in one ear, but they also fuse with the remainder of the syllable in the other ear to produce a percept equivalent to the original syllable (Rand, 1974; Cutting, 1976). This "duplex perception" demonstrates the simultaneous use of speech and nonspeech modes of perception and has recently been further explored in experiments that will be reviewed later in this paper.

Other authors have noted striking differences in subjects' responses depending on whether identical or similar stimuli were perceived as speech or nonspeech. For example, House, Stevens, Sandel, and Arnold (1962) found that an ensemble of speech stimuli was easier to learn than various ensembles of speechlike stimuli that, however, were not perceived as speech by the subjects (cf. also Grunke & Pisoni, Note 1). Several studies of categorical perception have shown that speech stimuli from a synthetic continuum are discriminated well across a phonetic category boundary, while nonspeech analogs or components of the same stimuli are discriminated poorly or at chance (e.g., Liberman, Harris, Eimas, Lisker, & Bastian, 1961; Liberman, Harris, Kinney, & Lane, 1961; Mattingly, Liberman, Syrdal, & Halwes, 1971). As long as two decades ago, House et al. (1962) concluded that "an understanding of speech perception cannot be achieved through experiments that study classical psychophysical responses to complex acoustic stimuli. ... Although speech stimuli are accepted by the peripheral auditory mechanism, their interpretation as linguistic events transfers their processing to some nonperipheral center where the detailed characteristics of the peripheral analysis are irrelevant" (p. 142). This conclusion is still valid, as the remainder of this paper will attempt to show.

Summary

Of the various paradigms reviewed by Cutting (1978) and Schouten (1980), some failed to support the existence of a speech mode of perception because they were irrelevant to begin with. As far as categorical perception and hemispheric specialization are concerned, some of the evidence may have been misinterpreted. The fact that categorical perception and left-hemisphere superiority can be obtained for certain nonspeech stimuli does away with earlier claims that these phenomena are speech-specific. However, it does not necessarily imply that similar patterns of results occur for the same reason in speech and nonspeech; and if they do, it is not necessarily true that the processes involved in the perception of nonspeech are more basic than, or the prerequisites for, those supporting speech perception. We have seen that there are other findings, not considered by Cutting and Schouten, that suggest that speech perception differs from nonspeech auditory perception. It must be acknowledged, however, that the empirical results are complex, and while they hardly argue against the existence of a speech mode, they do not provide an overwhelming amount of positive evidence either.

Certainly, the argument that speech perception is special would be strengthened if new, less controversial results could be brought to bear on the issue. The second part of this paper focuses on a set of rather recent findings that add a new dimension to the argument. Since these results are recent and have not been reviewed previously, they will be treated in more detail. They may be grouped into three categories: phonetic trading relations, context effects, and other perceptual integration phenomena. What is common to all of them is that they deal with *integration* (over frequency, time, or space) in phonetic perception.

THE NEW EVIDENCE

The Distinction Between Trading Relations and Context Effects

It is known from many previous studies that virtually every phonetic contrast is cued by several distinct acoustic properties of the speech signal. It follows that, within limits set by the relative perceptual weights and by the ranges of effectiveness of these cues, a change in the setting of one cue (which, by itself, would have led to a change in the phonetic percept) can be offset by an opposed change in the setting of another cue so as to maintain the original phonetic percept. This is a phonetic trading relation. According to Fitch, Halwes, Erickson, & Liberman (1980), there is a *phonetic equivalence* between two cues that trade with each other. I prefer to use this term in a slightly different way, for neither cue is perceived in isolation; rather, they are perceived together and integrated into a unitary phonetic percept. Therefore, the equivalence holds not so much between (a-b) units of Cue 1 and (c-d) units of Cue 2, but rather between the phonetic percept caused by setting a of Cue 1 and setting d of Cue 2 and the phonetic percept caused by setting b of Cue 1 and setting c of Cue 2. These two percepts are phonetically equivalent in the sense that they yield exactly the same distribution of identification responses and are difficult to discriminate (see below).

Trading relations occur among different cues for the same phonetic contrast. However, when the perception of a phonetic distinction is affected by preceding or following context that is not part of the set of direct cues for the distinction (as illustrated in the next paragraph), we speak of a context effect. The context may be "close," i.e., it may constitute portions of the same coherent speech signal; or it may be "remote," referring to the relation between separate stimuli in a sequence, or between a precursor and a test stimulus. (Of course, the distinction between close and remote context is, to some extent, arbitrary.) Effects of close context, which are of special interest to us, are similar to trading relations in that they can be cancelled by an appropriate change in one or another cue relevant to the critical phonetic distinction. Conversely, a trading relation could be described (inappropriately) as a context effect, with one cue (the context) affecting the perception of another (the target). Formally, trading relations and context effects are quite similar, but it is useful to distinguish them on theoretical grounds. The distinction is best illustrated with an example.

Mann and Repp (1980) presented listeners with fricative noises from a synthetic [ʃ]-[s] continuum, immediately followed by one of four periodic stimuli. The periodic stimuli derived from natural utterances of [ʃɑ], [sɑ], [ʃu], and [su], from which the fricative noise portion had been removed; thus, they contained formant transitions appropriate for either [ʃ] or [s], and the identity of the vowel was either [ɑ] or [u]. The results showed that, for a given ambiguous noise stimulus, listeners reported more instances of "s" when the following formant transitions were appropriate for [s] rather than [ʃ], and they also reported more instances of "s" when [u] followed rather than [ɑ]. The first effect is a trading relation, the second a context effect. The effect of formant transitions on perception of the [ʃ]-[s] distinction is a trading relation because the transitions are a cue to fricative place of articulation. They are also a direct consequence of fricative production, and this is obviously the reason why they are a cue to fricative perception. Note that the transitions are integrated with the fricative noise cue into a unitary phonetic percept; listeners do not perceive a noise plus transitions, or a fricative consonant followed by a stop consonant, although a stop would be perceived if the fricative noise were removed or silence were inserted between it and the periodic portion (Cole & Scott, 1973; Mann & Repp, 1980). The effect of vowel identity on fricative perception is different. Whether the vowel is [ɑ] or [u] is not a consequence of fricative production, and vowel quality therefore does not constitute a direct cue for fricative perception. The vowel is not perceptually integrated with the noise cue--it remains audible as a separate phonetic segment. It is appropriate here to say that the perceived vowel quality modifies the perception or interpretation of the fricative cues. This is a context effect, as distinct from a trading relation.[1]

As we will see below, trading relations and context effects have distinct (though related) explanations in a theory of phonetic perception, and it is that theoretical view that underlies the distinction in the first place. However, before we turn to the issue of explanation, a brief review of empirical findings shall be presented.

Phonetic Trading Relations

Overview

The fact that there are multiple cues for most phonetic contrasts has been known for a long time. Much of this early knowledge derives from the extensive explorations at Haskins Laboratories since the late 1940s. For example, Delattre, Liberman, Cooper, and Gerstman (1952) showed that the first two formants are important cues to vowel quality; Harris, Hoffman, Liberman, Delattre, and Cooper (1958) demonstrated that both second- and third-formant transitions contribute to the place-of-articulation distinction in stop consonants; and Gerstman (1957) found that both frication duration and rise-time are relevant to the fricative-affricate distinction. Lisker (1978b), drawing on observations collected over a number of years, listed no less than 16 distinguishable cues to the /b/-/p/ distinction in intervocalic position.

From these and many other studies, a nearly complete list of cues has been accumulated over the years. However, the data were typically collected by varying one cue at a time, although there are some exceptions, such as Hoffman's (1958) heroic study, which varied three cues to stop place of articulation simultaneously. Restrictions on the size of stimulus ensembles were imposed by the limited technology of the time, which made stimulus synthesis and test randomization very cumbersome. With the advent of modern computer-controlled synthesis and randomization routines, however, orthogonal variation of several cues in a single experiment became an easy task, and the limit to the number of stimuli was set by the patience of the listener rather than that of the investigator. The new technology led to a resurgence of interest in the way in which multiple cues cooperate in signalling a phonetic distinction. Since, for one reason or another, many of the early Haskins studies had remained unpublished, certain results that had been known for years by word of mouth or from preliminary reports only recently found their way into the literature, after having been replicated with contemporary methods.

A word is in order about the definition of cues. The traditional approach, exemplified especially by the Haskins work (including my own), has been to dissect a spectrographic representation of the speech signal, following essentially visual Gestalt principles. A cue, then, is a portion of the signal that can be isolated visually, that can be manipulated independently in a speech synthesizer constructed for that purpose, and that can be shown to have some perceptual effect. This way of defining cues has been challenged on two grounds: (1) The spectrogram is not the only, and not necessarily the best, representation of the speech signal. For example, the well-known work of Stevens and Blumstein (1978; Blumstein & Stevens, 1979, 1980) pursues the hypothesis that the shape of the total short-term spectrum at certain critical points in the signal constitutes a perceptual cue; thus, the individual formants and adjacent noise bursts are not treated as separate cues. Such a redefinition of cues is justified as long as it does not bypass the legitimate empirical issue of whether the elementary, spectrographically defined signal components are indeed integrated by the auditory system in this way (as they may be in the case of individual formants, but probably not in the case of other, more disparate types of cues). However, while definitions of such complex cues effectively combine information on one dimension (e.g., in the spectral domain), they typically sacrifice information on other dimensions (e.g., in the temporal domain). Thus, the onset spectra examined by Stevens and Blumstein are static and do not easily permit the description of dynamic change over time. The issue revolves, in large part, around the question how the perceptually salient information in the signal is best characterized--a question that, of course, lies at the heart of the present paper as well. The essential problem is that the totality of the cues for a given phonetic contrast apparently cannot be captured in a fully integrated fashion as long as purely physical (rather than articulatory or linguistic) terms are used.[2] (2) Another criticism of a more far-reaching sort denies altogether the usefulness of fractionating the speech signal into cues (see, e.g., Bailey & Summerfield, 1980). This view, which rests on the precepts of Gibsonian theory (Gibson, 1966), will be taken up in the concluding comments of this paper.

I will not attempt to review in detail all recent studies of phonetic trading relations, of which there are quite a few. A brief and selective overview shall suffice. Most studies had the purpose of clarifying the roles and surveying the effectiveness of different cues to various phonetic distinctions. Some studies that depart from this standard pattern will be considered later in more detail. Whereas the large majority of studies have used synthetic speech, some obtained similar information by cross-splicing components of natural utterances, or by combining such components with synthetic stimulus portions. Not all authors describe their findings as trading relations (a term used primarily by the Haskins group), but such relations are implied by the pattern of results.

Voicing cues. Many studies have investigated multiple cues to the voiced-voiceless distinction. For stop consonants in initial position, both voice onset time (VOT) and the first-formant (F1) transition contribute to the distinction (Stevens & Klatt, 1974; Lisker, Liberman, Erickson, Dechovitz, & Mandler, 1977). The critical feature of the F1 transition, which can be traded against VOT, is its onset frequency: If the onset frequency is lowered in a phonetically ambiguous stimulus, the VOT must be increased for a phonetically equivalent percept to obtain (Lisker, 1975; Summerfield & Haggard, 1977). Another cue that can be traded for VOT is the amplitude of the aspiration noise preceding the onset of voicing: If the amplitude of the noise is increased, its duration (i.e., the VOT) must be decreased to maintain phonetic equivalence (Repp, 1979). The fundamental frequency (F0) at the onset of the voiced stimulus portion is another relevant cue (Haggard, Ambler, & Callow, 1970) that presumably can be traded against VOT (see Repp, 1976, 1978b).

For stop consonants in intervocalic position, Lisker (1978b) has catalogued all the different aspects of the acoustic signal that contribute to the voicing distinction. They include the duration and offset characteristics of the preceding vocalic portion, the duration of the closure interval, the amplitude of voicing during the closure, and the onset characteristics of the following vocalic portion. Lisker's catalogue is based on a large number of studies, not all of which have been published; however, see Lisker (1957, 1978a, 1978c), Lisker and Price (1979), Price and Lisker (1979). Trading relations between voicing cues for intervocalic stops have also been studied in French (Serniclaes, 1974, Notes 2 & 3), and in German (Kohler, 1979).

The voicing distinction for stop consonants in final position has also been intensively studied. Here, the duration of the vocalic portion is important (especially if no release burst is present) as well as its offset characteristics, the properties of the release burst, and the duration of the preceding closure. Trading relations among these cues have been investigated by Raphael (1972, 1981), Wolf (1978), and Hogan and Rozsypal (1980), among others.

The voicing distinction for fricatives in initial position has been studied by Massaro and Cohen (1976, 1977) who focused on the trading relation between fricative noise duration and F0 at the onset of periodicity. In a similar fashion, Derr and Massaro (1980) and Soli (in press) studied the trading relations among duration of the periodic ("vowel") portion, duration of fricative noise, and F0 as cues to fricative voicing in utterance-final position. Earlier studies of these cues include Denes (1955) and Raphael (1972).

Place of articulation cues. Trading relations among place of articulation cues for stop consonants in initial position--F2 and F3 transitions, burst frequency and burst amplitude--were studied long ago by Harris et al. (1958) and Hoffman (1958), and more recently, by Dorman, Studdert-Kennedy, and Raphael (1977) and by Mattingly and Levitt (1980). For stop consonants in intervocalic position, Repp (1978a) found a trading relation between the formant transitions in and out of the closure, and Dorman and Raphael (1980) reported additional effects of closure duration and release burst frequency. Bailey and Summerfield (1980), in a series of painstaking experiments, investigated place cues for stops in fricative-stop-vowel syllables; these cues included the offset spectrum of the fricative noise, the duration of the closure period, and the formant frequencies at the onset of the vocalic portion. Repp and Mann (1981a) recently demonstrated a trading relation between fricative noise offset spectrum and vocalic formant transitions in similar stimuli. Fricative noise spectrum and vocalic formant transitions as joint cues to fricative place of articulation were investigated by Whalen (1981), Mann and Repp (1980), and Carden et al. (1981).

Manner cues. Cues to stop manner of articulation (i.e., to presence vs. absence of a stop consonant) following a fricative and preceding a vowel were investigated by Bailey and Summerfield (1980), Fitch et al. (1980), and Best, Morrongiello, and Robson (1981). In each case, the trading relation studied was that between closure duration and formant onset frequencies in the vocalic portion. The two last-named studies will be discussed in more detail below. Summerfield, Bailey, Seton, and Dorman (1981) have shown that duration and amplitude contour of the fricative noise preceding the silent closure also contribute to the stop manner contrast.

Several cues to the fricative-affricate distinction in initial position (rise-time, noise duration) were investigated by Gerstman (1957); see also van Heuven (1979). In a more recent set of experiments, Repp, Liberman, Eccardt, and Pesetsky (1978) traded vocalic offset spectrum, closure duration, and fricative noise duration as cues to a four-way distinction between vowel-fricative, vowel-stop-fricative, vowel-affricate, and vowel-stop-affricate. Trading relations among cues to the fricative-affricate distinction in final position were reported by Dorman, Raphael, and Liberman (1979: Exp. 5) and Dorman, Raphael, and Isenberg (1980).

Phonetic Equivalence

It is obvious that, whenever two or more cues contribute to a given phonetic distinction, they can be traded against each other, within certain limits. What is not so obvious is that two stimuli with equal response distributions are truly equivalent in perception. Since most data on trading relations were collected in identification tasks with a restricted set of response categories, subjects may have had no opportunity to report that certain stimuli sounded like neither of the alternatives. At a more subtle level, it may be the case that phonetically equivalent stimuli, even though they are labeled similarly, sound different in some way that subjects cannot easily explain in words. One way to assess this possibility is by means of a discrimination task.3

This was undertaken by Fitch et al. (1980) for the trading relation between silent closure duration and vocalic formant transition onsets as cues to stop manner in the "slit"-"split" distinction, and by Best et al. (1981) for the similar trading relation between silent closure duration and F1 transition onset in the "say"-"stay" contrast. First, these authors determined in an identification task how much silence was needed to compensate for a certain difference in formant onset frequency. Then they devised a discrimination task containing three different types of trials: On single-cue trials, the stimuli to be discriminated differed only in the spectral cue (formant onset frequency); they had the same setting of the temporal cue (silence). On cooperating-cues trials, the stimuli differed in both cues, such that the stimulus with the lower formant onsets (which favor "split" or "stay" percepts) also had the longer silence (which also favors "split" or "stay" percepts). On conflicting-cues trials, the stimuli again differed in both cues, but now the stimulus with the lower formant onsets had the shorter silence, so that one cue favored "split" ("stay") and the other "slit" ("say"). Since the silence difference chosen was the one found to compensate exactly for the spectral difference in the identification task, the stimuli in the conflicting-cues condition were (on the average) phonetically equivalent.[4]

The results of these experiments showed a clear difference among the three conditions: Subjects' discrimination performance in the category boundary region was best in the cooperating-cues condition, worst in the conflicting-cues condition, and intermediate in the single-cue condition. Thus, it is true that (approximately) phonetically equivalent stimuli, namely those in the conflicting-cues condition, are difficult to discriminate; they "sound the same," whereas stimuli in the cooperating-cues condition sound different, even though they exhibit the same physical differences on the two relevant dimensions. The pattern of discrimination results follows that predicted from identification data, showing that stimuli differing on two auditory dimensions simultaneously are still categorically perceived (given that perception is categorical when each of these dimensions is varied separately). It is likely that listeners could be trained to become more sensitive to the physical differences that do exist between phonetically equivalent stimuli, and the interesting question arises whether discrimination on cooperating-cues trials would continue t^ be superior to that on conflicting-cues trials. So far, no study has taken this approach. However, preliminary results from a related series of experiments (Repp, 1981b) indicate that some trading relations disappear when listeners try to discriminate pairs of stimuli that unambiguously belong to the same phonetic category (i.e., phonetically equivalent stimuli that are not from the boundary region), suggesting that these trading relations operate only when the stimuli are phonetically ambiguous. This leads us to the question of the origin of trading relations.

Explanation of Trading Relations: Phonetic or Auditory?

The large number of trading relations surveyed above poses formidable problems for anyone who would like to explain speech perception in purely auditory terms. Why should cues as diverse as, say, VOT and F1 onset, or silence and fricative noise duration, trade in the way they do? Auditory theory has only two avenues open: Either the cues are integrated into a unitary auditory percept at an early stage in perception (the auditory integration hypothesis), or selective attention is directed to one of the cues (which then must be postulated to be the essential cue for the relevant phonetic contrast), and the perception of that cue is affected by the settings of other cues (the auditory interaction hypothesis).

The auditory integration hypothesis is implicit in the work of Stevens and Blumstein (1978; Blumstein & Stevens, 1979, 1980). To account for the fact that release burst spectrum and formant transition onset frequencies are joint cues to place of articulation of syllable-initial stop consonants, Stevens and Blumstein assume that the perceptually relevant variable is the integrated spectrum of the first 25 msec or so of a stimulus. In other words, the burst (which is usually shorter than 25 msec) and the onsets of the several formant transitions are considered an integral auditory variable. Since both cues are spectral in nature and occur within a short time period, this is not an unreasonable hypothesis, notwithstanding the different sources of excitation (noise vs. periodic) of the two sets of cues in voiced stops.

In fact, Ganong (1978) found support for the perceptual integrality of burst and formant transition cues in an ingenious experiment involving interaural transfer of selective-adaptation effects. However, Stevens and Blumstein have had only limited success with automatic classification of stop consonants according to onset spectrum alone, and Kewley-Port (1981) recently demonstrated that automatic stop consonant identification can be improved by incorporating a measure of spectral change. Thus, even though onset spectrum may be an important cue, it does not contain all the relevant information in the signal.

The main problem with the auditory integration hypothesis seems to be that it applies only when the relevant cues are both spectral in nature, are of short duration, and occur simultaneously or in close succession. However, the cues are often spread out over a considerable stretch of time. For example, an explanation of the fact that both the formant transitions into and out of a stop closure contribute to the perceived place of articulation of a stop in medial position (Dorman & Raphael, 1980; Repp, 1978a; Repp & Mann, 1981a) would require integration of spectra across a closure, i.e., over as much as 100 msec. Such a long integration period seems unlikely; certainly, it is much longer than that envisioned by Stevens and Blumstein (1978). Trading relations that involve spectral and temporal cues (e.g., F1 onset and VOT for stop voicing in initial position) cannot be easily translated into purely spectral terms; and trading relations between purely temporal cues (e.g., silent closure duration and fricative noise duration for the fricative-affricate distinction in medial position) require a different explanation altogether. To be sure, there are some trading relations that do suggest auditory integration, such as that between VOT (i.e., aspiration noise duration) and aspiration noise amplitude (Repp, 1979), which is reminiscent of certain time-intensity reciprocities at the auditory threshold. In fact, preliminary data (Repp, 1981b) support this suggestion by showing that this trading relation operates independently of whether a listener is making phonetic or auditory judgments of speech stimuli. In other cases, however, the cues that participate in a trading relation are simply too diverse or too widely spread out to make auditory integration seem plausible. Or, to put it somewhat differently, whereas any such trading relation could be described as resulting from auditory integration, this integration would no longer seem to be motivated by general principles of auditory perception; thus, it would have to be considered a speech-specific process.

The auditory interaction hypothesis, which postulates that trading relations arise because perception of a primary cue is affected by other cues, has even less concrete evidence in its favor, in part because most of the relevant studies remain to be done. In particular, it is not clear whether auditory interactions (masking, contrast, etc.) of the kind and extent required to explain certain trading relations are at all plausible. For example, to explain the trading relation between VOT and F1 onset frequency as cues to stop consonant voicing, it would have to be the case that a noise-filled interval (VOT) sounds subjectively longer when followed by a periodic stimulus with a relatively low onset frequency. At present, there are no psychoacoustic data to support this hypothesis. Auditory psychophysics involving nonspeech stimuli of the degree of complexity of speech is still in its infancy (cf. Pastore, 1981). Perhaps, as more is learned about the perception of complex sounds and sound sequences, some auditory explanations of what now appear to be phonetic phenomena will be forthcoming.[5] One serious problem that has vexed researchers since the time of the early Haskins research is that of finding appropriate nonspeech analogs for speech stimuli. If the analogs are too similar to speech, they may be perceived as speech and thereby cease to be good analogs and become bad speech. If they are too different from speech, the generalizability of the findings to speech may be questioned. There is a way out of this dilemma: If stimuli could be constructed that are sufficiently like speech to be perceived as speech by some listeners but not by others (perhaps prompted by different instructions), or even by the same listeners on different occasions, and if different results are obtained in the two conditions (e.g., two cues trade in one but not in the other), this would then be proof of specialized perceptual processes serving speech perception.

It is from this perspective that a recent study by Best et al. (1981) receives special importance. These authors investigated the trading relation between silent closure duration and F1 transition onset frequency as cues to stop manner in the "say"-"stay" contrast. After replicating the results obtained with the similar "slit"-"split" contrast by Fitch et al. (1980), they proceeded to test for the presence of a similar trading relation in "sinewave analogs" of the synthetic "say"-"stay" stimuli. Sinewave analogs are obtained by imitating the formant trajectories of (voiced) speech stimuli with pure tones. Such analogs of simple CV syllables have been used previously by Cutting (1974) and by Bailey, Summerfield, and Dorman (1977), whose work is discussed below; recently, Remez, Rubin, Pisoni, and Carrell (1981) successfully synthesized whole English sentences in that way. The interesting thing about these stimuli is that they are heard as nonspeech whistles by the majority of naive listeners, but they may be heard as speech when instructions point out their speechlikeness, or spontaneously after prolonged listening. Once heard as speech, it is difficult (if not impossible) to hear them as pure whistles again, although the speech heard retains a highly artificial quality (Remez et al., 1981). This phenomenon was exploited by Best et al. in their main experiment.

They constructed sinewave analogs of a "say"-"stay" continuum by following a noise resembling [s]-frication with varying periods of silence and a sine-wave portion whose component tones imitated the first three formants of the periodic portion of the speech stimuli. There were two versions of the sinewave portion, one with a low onset of the tone simulating F1, and one with a high onset. (In speech stimuli, less silence is needed to change "say" to "stay" when F1 has a low onset than when it has a high onset.) The sinewave stimuli were presented to listeners in an AXB format, where the critical X stimulus had to be designated as being more similar to either the A or the B stimulus, which were analogs of a clear "say" (no silence, high F1 onset) and a clear "stay" (long silence, low F1 onset), respectively. Some of the subjects were told that the stimuli were intended to sound like "say" or "stay," whereas others were only told that the stimuli were computer sounds. After the experiment, the subjects were divided into those who reported that they heard the stimuli as "say"-"stay," either spontaneously or after instructions, and into those who reported various auditory impressions or inappropriate speech percepts. Only members of the first group, who--according to their self-reports--employed a phonetic mode of perception, showed a trading relation between silence and F1 onset frequency, and this trading relation resembled that obtained with synthetic speech stimuli. None of the other subjects showed this pattern of results. These other subjects could be further subdivided into two groups: those who reported that the stimuli differed in the amount of separation between the two stimulus portions (noise and sinewaves), and those who reported that the stimuli differed in the quality of the onset of the second portion ("water dripping," "thud," etc.). The AXB results substantiated these reports: The results of the first group indicated that the subjects paid attention only to the silence cue, whereas the second group seemed to make their judgments primarily on the basis of the spectral cue (F1-analog onset frequency). The response patterns of the two groups were radically different from each other, and both were different from the group who heard the stimuli as speech. It seems reasonable to conclude that the subjects in the former two groups employed an auditory mode of perception. Being in this mode, they were unable to integrate the two cues into a unitary percept and instead focused on one or the other cue separately, thereby disconfirming the auditory integration hypothesis for this set of cues.[6] There was some evidence of an auditory interaction in that those listeners who paid attention to the spectral cue were affected by the setting of the temporal cue. However, this effect was not sufficiently strong to account for the trading relation observed in speech-mode listeners; moreover, those subjects who focused on the silence cue (which is the primary cue for stop manner) were not affected at all by the setting of the spectral cue.

The results of Best et al. provide the strongest evidence we have so far that a trading relation is specific to phonetic perception: When listeners are not in the speech mode, the trading relation disappears and selective attention to individual acoustic cues becomes possible. The data argue against any auditory explanation of the trading relation at hand, and they support the existence of a phonetic mode of perception that is characterized by specialized ways of stimulus processing. Results from a recent study (Repp, 1981b) further confirm the phonetic nature of the trading relation between silence and F1 onset for the "say"-"stay" distinction by showing that it is obtained only in the phonetic boundary region of the speech continuum (i.e., when listeners can make a phonetic distinction) but not within the "stay" category (i.e., when listeners cannot make a phonetic distinction and must rely on auditory criteria for discrimination). We may suspect that many other trading relations will behave similarly. This is already indicated for the trading relation between closure duration and fricative noise duration in the "say shop"-"say chop" distinction (Repp, 1981b) and for that between fricative noise spectrum and formant transitions in the [ʃ]-[s] distinction (Repp, 1981a, discussed in the next section).

How, then, are trading relations to be explained, if not in terms of auditory interactions or integration? The proposed answer is this: Speech is produced by a vocal tract, and the production of a phonetic segment (assuming that such segments exist at some level in the articulatory plan) has complex and temporally distributed acoustic consequences. Therefore, the information supporting the perception of the same phonetic segment is acoustically diverse and spread out over time. The perceiver recovers the abstract units of speech by integrating the multiple cues that result from their production. The basis for that perceptual integration may be conceptualized in two ways. One is to state that listeners know from experience how a given phonetic segment "ought to sound like" in a given context. Since phonetic contrasts almost always involve more than one acoustic property, trading relations among these properties must result when the stimulus is ambiguous because, in this view, it is being evaluated with reference to idealized representations or "prototypes" that differ on all these dimensions simultaneously: A change in one dimension can be offset by a change in another dimension, so that the perceptual distances from the prototypes remain constant. The other possibility is that perceptual integration does not require specific knowledge of speech patterns (whose form of memory storage is difficult to conceptualize) but is predicated directly upon the articulatory information in the signal. In other words, trading relations may occur because listeners perceive speech in terms of the underlying articulation, and inconsistencies in the acoustic information are resolved to yield perception of the most plausible articulatory act. This explanation thus requires that the listener have at least a general model of human vocal tracts and of their ways of action. The question remains: How much must an organism know about speech to exhibit a phonetic trading relation? An important issue for future research will be the question whether phonetic trading relations are obtained in human infants, and if not, how and when they begin to develop.[7]

Context Effects

Effects Due to Immediate Phonetic Context

Like phonetic trading relations, certain kinds of phonetic context effects have been known for a long time. The most familiar example is, perhaps, the dependence of stop release burst perception on the following vowel. Liberman, Delattre, and Cooper (1952) showed that, when noise bursts of varying frequencies are followed by different steady-state periodic stimuli, the stop consonant categories reported by listeners may depend on the quality of the vowel. For example, if a noise burst centered at 1600 Hz is followed by steady states appropriate for [i] or [u], listeners report "p," but if [a] follows, they report "k."

A similar effect has been reported by Summerfield (1975) who found that the nature of the vowel influences the location of the boundary on a continuum of stop-consonant-vowel syllables varying in VOT. This context effect may actually be a trading relation because it probably reflects the influence of F1 onset (rather than vowel quality _per se_) on the voicing decision, i.e., a trading relation between F1 onset and VOT (cf. Summerfield & Haggard, 1974, 1977). Recently, Summerfield (in press) conducted an important series of experiments in which he tested whether this effect has an auditory basis. He used speech stimuli varying in VOT and in the F1 frequency of the following steady-state vocalic portion, and he compared their perception with that of two kinds of nonspeech analogs. One was a tone-onset-time (TOT) continuum (Pisoni, 1977) that varied the relative onset time of two pure tones of fixed frequency, matched in frequency and amplitude to the first two formants of the speech stimuli. The frequency of the lower tone was varied to simulate different F1 onset frequencies. The other set of nonspeech stimuli formed a noise-onset-time (NOT) continuum (cf. Miller et al., 1976) that varied the lead time of a noise-excited steady-state F2 relative to a periodically excited steady-state F1. Different F1 onset frequencies were simulated by varying the frequency of F1. The stimuli were presented for identification as "g" or "k" (speech) or as "simultaneous onset" vs. "successive onset" (nonspeech). While the VOT boundary exhibited the expected sensitivity to F1 onset frequency, neither nonspeech continuum evinced any reliable influence of F1(-analog) frequency on listeners' judgments. Pastore et al. (1981) recently reported a similar failure to find equivalent effects of two different secondary variables (rise time and trailing stimuli) on VOT and TOT category boundaries. These results suggest that the context effect obtained in speech does not have an auditory basis but is specific to the phonetic mode. (However, see Footnote 7.)

An effect of vocalic context on the perception of stop consonant place of articulation was investigated by Bailey et al. (1977). These authors constructed two synthetic speech continua ranging from [b]+vowel to [d]+vowel by varying the transition onset frequencies of F2 and F3. The two continua differed in the terminal (steady-state) frequency of F2, which was high in one and low in the other. On each continuum, the transition onsets were arranged so that the center stimulus had completely flat F2 and F3, while both transitions rose in one endpoint stimulus to the same degree as they fell in the other endpoint stimulus. When these stimuli were presented to subjects for classification in an AXB task, it turned out that the category boundaries were at different locations on the two continua, neither being exactly in the center: one (on the continuum with the low-F2 vowel) was displaced toward the [d] end, while the other boundary was displaced toward the [b] end. Bailey et al. wished to test whether this difference (a kind of context effect, especially when "rising vs. falling transitions" is considered the relevant cue, rather than absolute transition onset frequency, which varied with context) has a psychoacoustic basis. They pioneered in using sinewave analogs for that purpose. The sinewave stimuli were presented in the same AXB paradigm to a group of subjects that was subdivided afterwards according to self-reports whether or not the stimuli were heard as speech. It turned out that those listeners who claimed to hear [b] and [d] had their category boundaries on the two continua at different locations that corresponded to those found with speech stimuli. The other listeners, however, who reported *only nonspeech impressions*, had their boundaries close to the centers of both continua, as one might predict on psychophysical grounds. This experiment provided evidence that phonetic categorization is based on principles different from those of auditory psychophysics. Presumably—although this was not shown directly by Bailey et al.—the asymmetrical boundaries obtained with speech stimuli were in accord with the acoustical characteristics of typical stop consonants in these particular vocalic contexts.

Let us turn now to other context effects that are of special interest because they involve segments not as obviously interdependent as stop consonants and following vowels. One effect concerns the influence of vocalic context on fricative perception. If a noise portion ambiguous between [ʃ] and [s] is followed by a periodic portion appropriate for a rounded vowel such as [u], listeners are more likely to report "s" than if the following vowel is unrounded, e.g., [a] (Kunisaki & Fujisaki, Note 5; Mann & Repp, 1980; Whalen, 1981). A preceding vowel has a similar, but smaller effect (Hasegawa, 1976). In addition to roundedness, other features of the vowel (such as the front-back dimension) also seem to play a role (Whalen, 1981). Repp and Mann (1981a) also discovered a small but reliable effect of a following stop consonant on fricative perception: Listeners are more likely to report "s" when the formant transitions in the following vocalic portion (separated from the noise by a silent closure interval) are appropriate for [k] than when they are appropriate for [t].

Several effects of context on the perception of stop consonants have been discovered in recent experiments. Mann and Repp (1980) found that, in fricative-stop-vowel stimuli, listeners are more likely to report "k" when vocalic stimuli with formant transitions ambiguous between [t] and [k] are preceded by an [s]-noise plus silence than when they are preceded by an [ʃ]-noise plus silence. They showed that the effect has two components, one due to the spectral characteristics of the fricative noise (perhaps an auditory effect) and the other to the category label assigned to the fricative (which must be a phonetic effect). Subsequently, Repp and Mann (1981a) showed the context effect to be independent of the effect of direct cues to stop place of articulation in the fricative noise offset spectrum (which proves that it is a true context effect and not a trading relation), and they also ruled out simple response bias as a possible cause. In a further experiment, Mann (1980) found that, when stimuli ambiguous between [da] and [ga] were preceded by either [al] or [ar], listeners reported many more "g" percepts after [al] than after [ar]. In experiments with vowel-stop-stop-vowel stimuli, Repp (1978a, 1980a, 1980b) found various perceptual interdependences between the two stops cued by the formant transitions on either side of the closure interval; in particular, perception of the first stop was influenced strongly by the second.

How are all these effects to be explained? Auditory explanations would have to be formulated in the manner of the interaction hypothesis for trading relations: The perception of the relevant acoustic cues is somehow affected by the context. As in the case of trading relations, however, no plausible mechanisms that might mediate such effects have been suggested, and no similar effects with nonspeech analogs have been reported so far. On the other hand, reference to speech production provides a straightforward explanation of most, if not all, context effects. Just as trading relations reflect the dynamic nature of articulation (of a given phonetic segment), so are context effects accounted for by <u>coarticulation</u> (of different phonetic segments). The articulatory movements characteristic of a given phonetic segment exhibit contextual variations that may be either part of the articulatory plan (allophonic variation, or anticipatory coarticulation) or due to the inertia of the articulators (perseverative coarticulation). Presumably, human listeners possess implicit knowledge of this coarticulatory variation.

Coarticulatory effects corresponding to the perceptual phenomena just cited have been observed in most cases. Thus, it is well known that the release burst spectrum of stop consonants varies with the following vowel (Zue, Note 6) in a manner quite parallel to the perceptual findings of Liberman et al. (1952). Fricative noises exhibit a downward shift in spectrum when they precede or follow a rounded vowel, due to anticipatory or carry-over lip rounding (Fujisaki & Kunisaki, 1978; Hasegawa, 1976; Mann & Repp, 1980), which explains the effect of vocalic context on fricative perception. The formant transitions of stop consonants vary with preceding fricatives (Repp & Mann, 1981a, 1981b) and liquids (Mann, 1980) in a manner consistent with the corresponding perceptual effects. Thus, the available evidence suggests that most perceptual context effects are parallelled by coarticulatory effects. The implication is, then, that listeners expect coarticulation to occur and compensate for its absence in experimental stimuli by shifting their response criteria accordingly. For example, if an [ʃ]-like noise followed by [u] is not sufficiently low on the spectral scale (as it should be because of anticipatory lip rounding), it might be perceived as an "s." Thus, the evidence is highly persuasive that context effects, just like trading relations, reflect the listeners' intrinsic knowledge of articulatory dynamics.

A critical test of the auditory vs. phonetic explanations of context effects can again be performed with appropriate nonspeech analogs, or with stimuli that can be perceived as either speech or nonspeech. Two such studies (Bailey et al., 1977; Summerfield, in press) were discussed above. In a recent experiment, I took an alternative approach (Repp, 1981a): Rather than using nonspeech stimuli that can be perceived as speech, I used speech stimuli (a portion of) which can be fairly readily perceived as nonspeech. Although it is usually difficult to abandon the phonetic mode when listening to speech, except in cases where the speech is strongly distorted or poorly synthesized, fricative-vowel syllables offer an opportunity to do so because they contain a sizable segment of fairly steady-state noise whose auditory properties ("pitch," length, loudness) are relatively accessible. In my study, the fricative noise spectrum was varied along a continuum from [ʃ]-like to [s]-like, and the vowel was either [a] or [u]. It was known from earlier experiments (Mann & Repp, 1980) that listeners are more likely to label the fricative "s" in the context of [u] than in the context of [a]. A secondary cue to the [ʃ]-[s] distinction was deliberately confounded with the context effect: The [a] vocalic portion contained formant transitions appropriate for [ʃ], and the [u] portion contained transitions appropriate for [s]; this increased the differential effect of the two vocalic contexts on fricative identification. (Thus, this experiment tested a context effect and a trading relation at the same time.) The stimuli were subsequently presented in a same-different discrimination task where the difference to be detected was in the spectrum of the noise portion, and the vowels were either the same or different, but irrelevant in any case. The majority of naive subjects perceived these stimuli fairly categorically: Their discrimination performance was poor; the pattern of responses suggested that they relied on category labels; and there were pronounced effects of vocalic context, just as in previous labeling tasks. Two subjects, however, performed much better than the others. Their data resembled those of three experienced listeners who also participated in the experiment. Comments and introspections of these subjects suggested that they were able to bypass or ignore phonetic categorization and to focus instead on the spectral properties (the "pitch") of the fricative noise. The crucial result was that these listeners not only performed much better than the rest (which supports the hypothesis that they employed an auditory mode of perception), but that they did not show any effect of vocalic context. These results were confirmed in a follow-up study where naive listeners were induced (with some success) to adopt an auditory listening strategy. These experiments demonstrate that vocalic context affected the perceived phonetic category of the fricative but not the perceived pitch quality of the noise. Therefore, the context effect due to the quality of the vowel, as well as the cue integration underlying the contribution of the vocalic formant transitions to fricative identification, must be phonetic in nature.

Speaker Normalization Effects

A phenomenon related to the context effects just discussed is that of speaker normalization. In an experimental demonstration of this effect, the perception of a critical phonetic segment is influenced, not by a phonetic change in an adjacent segment, but by an acoustic change such as might result from a change in speaker. For example, a (roughly proportional) upward shift of vowel formants on the frequency scale signifies that the speech signal originated in a smaller vocal tract. (How listeners "decide" that the same vowel has been produced by a smaller vocal tract, rather than a different vowel by the same vocal tract, is an unresolved issue.) Such a change may influence the perception of phonetic segments in the vicinity, as long as the listener perceives the whole test utterance as coming from a single speaker's vocal tract.

Although speaker normalization is a well-recognized problem in speech recognition research, there have been relatively few experimental studies. Rand (1971) constructed stop consonant continua ranging from /b/ to /d/ to /g/ by varying the onset of the F2 transition of three synthetic two-formant stimuli intended to represent, respectively, an /æ/ produced by a large vocal tract, an /æ/ produced by a small vocal tract (differing from the former only in F2 frequency), and an /ɛ/ produced by a large vocal tract (differing from the former only in F1 frequency). The results showed similar category boundaries (expressed in terms of absolute F2 onset frequency) for the two stimulus continua associated with large vocal tracts, but a shift towards higher frequencies on the continuum associated with a small vocal tract. Rand interpreted his findings as evidence for perceptual normalization, although this may not be the only possible explanation.

In a more recent study, May (1976) followed fricative noises from a synthetic [ʃ]-[s] continuum with one of two synthetic periodic portions, intended to represent the same vowel produced by two differently-sized vocal tracts. The [ʃ]-[s] boundary shifted as expected: Listeners reported more "s" percepts in the context of the larger vocal tract. Subsequently, Mann and Repp (1980) conducted a similar experiment in which synthetic fricative noises were followed by vocalic portions derived from natural utterances produced by a male or a female speaker. The results replicated those by May. These findings are consistent with the fact that smaller vocal tracts (females) produce fricative noises of higher average frequency than large vocal tracts (males) (Schwartz, 1968).

To these results must be added the evidence from studies that have shown speaker normalization effects due to "remote" context, i.e., due to other stimuli in a sequence or to precursor stimuli or phrases (e.g., Ladefoged & Broadbent, 1957; Strange, Verbrugge, Shankweiler, & Edman, 1976; Summerfield & Haggard, 1975). They all demonstrate the same point: Listeners interpret the speech signal in accordance with the perceived (or expected) dimensions of the vocal tract that produced it. Information about vocal tract size is picked up in parallel with information about articulator movements; these are, respectively, the static and dynamic (or structural and functional) aspects of articulatory information. Speaker normalization effects are difficult to explain in terms of a general auditory theory that does not make reference to the mechanisms of speech production. Although some effects could, in principle, result from auditory contrast, interactions of similar complexity have not yet been demonstrated in nonspeech contexts.

Rate Normalization Effects

The somewhat larger literature on perceptual effects of speaking rate has recently been thoroughly reviewed by Miller (1981). Rate normalization, like speaker normalization, is a kind of context effect, and it can be produced by either close or remote context. Rate normalization is said to occur when the perception of a phonetic distinction signalled by a temporal cue (i.e., by the duration of a stimulus portion, or by the rate of change in some acoustic parameter) is modified after a temporal change is introduced in portions of the context that are not themselves cues for the perception of the target segment.

Only a few representative findings shall be mentioned here. Miller and Liberman (1979) examined the stop-semivowel distinction (/ba/-/wa/), cued by the duration and rate of the initial formant transitions, and found that the category boundary shifted systematically with the duration of the vocalic portion (i.e., of the whole stimulus). A corresponding shift of the discrimination peak in an oddity task was reported by Miller (1980). This effect may have an auditory basis, for it has not only been found in human infants (Eimas & Miller, 1980) but also with analogous nonspeech stimuli (Carrell, Pisoni, & Gans, Note 7). However, it may also be argued that simple durational variation is not sufficient to create variations in perceived speaking rate.

Fitch (1981) recently attempted to dissociate information about speaking rate from phonetically distinctive durational variation. The phonetic distinction studied was that between [dabi] and [dapi], as cued by the duration of the first stimulus portion ([dab] or [dap]). By manipulating the duration of natural utterances produced at different rates, she was able to show that speaking rate had a perceptual effect separate from that of physical duration. Thus, the information about speaking rate seems to be carried, in part, by more complex structural variables, such as the rate of spectral change in the signal. Soli (in press) has recently obtained similar results in a thorough investigation of cues to the [jus]-[juz] distinction. These findings are considerably more difficult to explain by psychoacoustic principles.

The most convincing instances of rate normalization derive from studies that varied remote context. The perception of a variety of phonetic distinctions is sensitive to the perceived rate of articulation of a carrier sentence (e.g., Miller & Grosjean, 1981; Pickett & Decker, 1960; Summerfield, 1981). Miller and Grosjean (1981) showed that the articulation rate of the carrier sentence was more important than its pause rate, even though the critical phonetic contrast ("rabid"-"rapid") was cued primarily by the perceived duration of a silent interval. Findings such as these suggest that speaking rate is a rather abstract property whose perception requires an appreciation of articulatory and linguistic variables (cf. also Grosjean & Lane, 1976). Summerfield (1981) has shown that the rate of a nonspeech carrier (a melody) does not affect speech perception, confirming that the listener's rate estimate must derive from speech to be relevant.

These findings are just a sampling of a much larger literature on perceptual adjustments for speaking rate (see Miller, 1981). Whether or not there are corresponding contextual effects in the judgment of auditory duration is not known (except for the above-cited study by Carrell et al., Note 7), although there is some plausibility in the hypothesis that the durations of adjacent or corresponding auditory intervals are judged relative to each other. Perhaps because this hypothesis seems more plausible than possible auditory explanations of other context effects in speech, there have been few attempts so far to simulate speaking rate effects using nonspeech analog stimuli. However, there is some evidence that even simple durational changes may be interpreted differently in speech and nonspeech modes. Smith (1978) presented two identical syllables in succession and varied their relative durations. Listeners had to judge either which syllable was more stressed (a linguistic judgment) or which syllable was longer in duration (an auditory judgment). The two kinds of judgment diverged: Stress judgments exhibited a tendency for the first syllable to be judged stressed, whereas duration judgments showed no such bias. These results indicate that the linguistic function of acoustic segment duration cannot be directly predicted from auditory judgments of that duration. Presumably, in speech perception, acoustic segment duration is interpreted, as are all other cues, within a framework of tacitly known articulatory patterns and constraints, such as the well-known lengthening of a final syllable (Klatt, 1976).

Sequential (Remote) Context Effects

Context effects due to preceding and following stimuli in a test sequence are a ubiquitous phenomenon and well-known also in auditory psychophysics. They include effects of neighboring stimuli (preceding and/or following a target stimulus), as well as effects due to a whole series of preceding stimuli, referred to variously as selective adaptation, anchoring, range, or frequency effects. Even though these effects are clearly not in any way specific to speech--and speech stimuli are by no means immune to them, as was once believed with regard to anchoring (Sawusch & Pisoni, 1973; Sawusch, Pisoni, & Cutting, 1974)--the pattern of the data obtained for speech may nevertheless exhibit peculiarities not observed with nonspeech stimuli. The most striking of these is, of course, the relative stability of phonetic boundaries. Although all boundaries can be shifted to some extent by contextual influences, most boundaries do not change very much. (Isolated vowels are a significant exception--see below.) Presumably, this is so because listeners have internal criteria based on their long experience with speech, and especially with their native tongue. It might be argued that phonetic boundaries are stable because they coincide with auditory boundaries of some sort. However, the evidence for such a coincidence is not convincing (see my earlier discussion of categorical perception), and nonhuman subjects seem to exhibit much larger range-contingent boundary shifts for speech stimuli than adult human subjects (Waters & Wilson, 1976).

Another example of an interesting discrepancy between speech and nonspeech is provided by the pattern of vowel context effects. Repp et al. (1979) found not only that isolated synthetic vowel stimuli presented in pairs exhibit large contextual effects (as shown earlier by Fry, Abramson, Eimas, & Liberman, 1962; Lindner, 1966; Thompson & Hollien, 1970; and others), but also that backward contrast (the influence of the second stimulus on perception of the first) was stronger than forward contrast (the influence of the first stimulus on perception of the second). These results become interesting in the light of later findings that nonspeech stimuli show (surprisingly) much smaller contrast effects than isolated vowels and no (or the opposite) difference between forward and backward contrast. Healy and Repp (in press) obtained these results by comparing vowels from an [i]-[I] continuum with brief nonspeech "timbre" stimuli (single-formant resonances of varying frequency, labeled as "low" or "high"). Fujisaki and Shigeno (1979) also compared vowels with timbre stimuli that, however, had the same duration, and still found a large difference in the magnitude of contrast effects, and larger backward than forward contrast for vowels only. Shigeno and Fujisaki (Note 8) compared phonetic category judgments of vowels varying in spectrum with pitch judgments of a single vowel varying in F0. While the former condition replicated earlier findings (large contrast effects, more backward than forward contrast), there were no contrast effects at all in the latter condition. While it seems possible that an auditory explanation of these results will eventually be found, the peculiar flexibility of vowel perception may also be grounded in the special status of vowels as nuclear elements in the speech message. Perhaps, the modifiability of vowel perception corresponds to the remarkable contextual variability vowels exhibit in the speech signal.

Other Perceptual Integration Effects

A discussion of evidence for a phonetic mode of perception would not be complete without mention of two strands of research that make a particularly important contribution. They both deal with the integration of cues separated not in time but in space or even occurring in different modalities.

Duplex Perception

Duplex perception is the newly coined (Liberman, 1979) name for a phenomenon originally discovered by Rand (1974) and described earlier in this paper: An isolated formant transition presented to one ear simultaneously with the "base" (a synthetic CV syllable bereft of that formant transition) in the other ear is perceived as a lateralized nonspeech "chirp" although, at the same time, it contributes (presumably, by some process of central integration) to the perception of the syllable in the other ear. The phenomenon by itself shows that the same input may be perceived in auditory and phonetic modes at the same time: the transition is auditorily segregated, yet phonetically integrated with the base. Several recent studies show that various experimental variables affect either the auditory or the phonetic part of the duplex percept, but not both.

Thus, Isenberg and Liberman (1978) varied the intensity of the isolated transition. The subjects perceived changes in the loudness of the chirp, but they could not detect any change in the loudness of the syllable in the other ear, even though they perceived the phonetic segment specified by the transition. Liberman, Isenberg, and Rakerd (1981) immediately preceded the base with a fricative noise appropriate for [s], which (in the absence of any intervening silence) inhibited the perception of the stop consonant ([p] or [t]) that the base in conjunction with the transition in the other ear otherwise would have generated. Listeners found it difficult to discriminate [s]+[pa] and [s]+[ta] as long as they attended to the side on which the speech was heard, for both stimuli sounded like [sa]. However, their discrimination of [p]-chirps from [t]-chirps in the other ear was highly accurate. Recently, Mann, Madden, Russell, and Liberman (1981) used the duplex perception paradigm to examine further the effect (discovered by Mann, 1980) of a preceding liquid on stop consonant perception. When the syllables [al] or [ar] preceded the base of a stimulus from a [ta]-[ka] continuum, the context effect was obtained in phonetic perception (more [ka] percepts following [al]) while the perception of the isolated transition in the other ear was unaltered.

Effects similar to duplex perception have been reported, where some nonspeech stimulus in one ear affected phonetic perception in the other ear while retaining its nonspeech quality. For example, Pastore (1978) found that, when the syllable [pa] in one ear was accompanied by a burst of noise in the other ear, phonetic perception changed to [ta]. Apparently, the noise--even though it did not have the appropriate timing, duration, and envelope--was interpreted by listeners as a [t]-release burst and was integrated with the syllable in the other ear. There is no doubt, however, that listeners nevertheless continued to hear a nonspeech sound in the ear in which the noise occurred. The finding of Repp (1976) that the pitch of an isolated vowel in one ear affects the perception of the voiced-voiceless distinction for stop-consonant-vowel syllables in the other ear may be taken as another instance of duplex perception. Presumably, listeners could have accurately judged the pitch of the isolated vowel without destroying its phonetic effect.

Duplex perception phenomena provide evidence for the distinction between auditory and phonetic modes of perception. They show that the auditory mode can gain access to the input from individual ears while the phonetic mode, under certain conditions, operates on the combined input from both ears. The "phonological fusion" discovered by Day (1968)--two dichotic utterances such as "banket" and "lanket" yield the percept "blanket"--is yet another example of the abstract, nonauditory level of integration that characterizes the phonetic mode.

Audio-Visual Integration

Perhaps the most important recent discovery in the field is the finding of an influence of visual articulatory information on phonetic perception (McGurk & MacDonald, 1976; MacDonald & McGurk, 1978; Summerfield, 1979). Of course, it has been known for a long time that lip reading aids speech perception, especially for the hard of hearing, but only recently has it become clear how tight audio-visual integration can be. McGurk and MacDonald (1976) presented a video display of a person's face saying simple CV syllables in synchrony with acoustic recordings of syllables from the same set. When the visual and auditory information disagreed, the visual information exerted a strong influence on the subjects' percepts, primarily due to the readily perceived presence vs. absence of visible lip closure. Thus, when a visual /da/ or /ga/ was paired with an auditory /ba/, subjects usually reported /da/.8

The interpretation of this finding is straightforward and of great theoretical significance. Clearly, subjects somehow combine the articulatory information gained from the visual display with that gained from the acoustic signal. In Summerfield's (1979) words, "optical and acoustic displays are co-perceived in a common metric closely related to that of articulatory dynamics" (p. 314). This phenomenon provides some of the strongest evidence we have for the existence of a speech-specific mode of perception that makes use of articulatory, as opposed to general auditory, information. The common metric of visual and auditory speech input represents a modality-independent, presumably articulation-based level of abstraction that is the likely site of the integration and context effects reviewed above. Phonetic perception in the auditory modality (when speech enters through the ears) is likely to be in every sense as abstract as it is in the visual modality (when articulatory movements are observed directly).

In a recent ingenious study, Roberts and Summerfield (1981) used the audio-visual technique to demonstrate that selective adaptation of phonetic judgments is a purely auditory effect. Although conflicting visual information changed the listeners' phonetic interpretation of an adapting stimulus, it had no effect whatsoever on the direction or magnitude of the adaptation effect. Besides its implications for the selective adaptation paradigm (cf. also Sawusch & Jusczyk, 1981), this elegant study provides further evidence for the autonomy of phonetic perception.

Disruption of Perceptual Integration

As was pointed out in the discussion of speaker normalization effects, a simulated change in vocal tract size (or in any other speaker characteristic, such as fundamental frequency) must not disrupt the perceptual coherence of an utterance if a normalization effect shall be observed. In the case of formant transitions leading into a vocalic stimulus portion, or of an aperiodic portion (fricative noise) being followed by a periodic portion, perceptual coherence is easily maintained when the formant frequencies of the vowel are changed. However, when two periodic signal portions appropriate to different vocal tracts are juxtaposed, a change in speaker may be perceived, and this may lead to the disruption of whatever perceptual interactions (trading relations or context effects) may have taken place between the two periodic signal portions. There are several examples of this phenomenon in the recent literature.

For example, Darwin and Bethell-Fox (1977) showed that, by changing fundamental frequency abruptly at points of transition, a speech stimulus originally perceived as a smooth alternation of a liquid consonant (or semivowel) and a vowel could be changed into a train of stop-vowel syllables perceived as being produced in alternation by two different speakers. The manipulation of F0 signalled a change in source and thus "split" the formant transitions into portions that effectively became new cues, signalling stop consonants rather than liquids or semivowels.

Dorman et al. (1979: Exp. 6) studied a situation in which the perception of a syllable-final stop consonant depends on whether or not there is a sufficient period of (near-)silence to indicate closure. An utterance such as /babda/ is generally perceived as /bada/ if the stop closure interval is removed. Dorman et al. found, however, that when the first syllable, /bab/, is produced by a male speaker and the second syllable, /da/, by a female speaker, the syllable-final stop in /bab/ is clearly perceived. Because of the perceived change in speakers, listeners no longer recognize the absence of a closure interval; the critical syllable-final stop is now in utterance-final position. Interestingly, two subjects who reported that they did not notice a change in speaker, also failed to perceive the syllable-final stop consonant in the absence of closure.

Conversely, an interval of silence in an utterance may lose its perceptual value when a change of speaker is perceived to occur across it (Dorman et al., 1979: Exp. 7): When silence is inserted into the utterance "say shop" immediately preceding the fricative noise, listeners report "say chop". However, when "say" is spoken by a male voice and "shop" by a female voice, this effect no longer occurs; the silence loses its phonetic significance, and the second syllable remains "shop."9

This effect was further investigated by Dechovitz, Rakerd, and Verbrugge (1980) who varied the perceived continuity of the test utterance "Let's go shop (chop)" by having speakers produce either the whole phrase or just "Let's go." Silence inserted (or removed from) between the "go" and the "shop (chop)" of a continuous utterance had the expected effect on phonetic perception: "shop" was perceived as "chop" when silence was present, and "chop" was perceived as "shop" when there was no silence. However, when the "Let's go" with phrase-final intonation was followed by either "shop" or "chop" from a different production, there were no such effects: "shop (chop)" remained "shop (chop)." Interestingly, these authors found that a change of speaker from female to male between "Let's go" and "shop (chop)" did not disrupt perceptual integration as long as the "Let's go" derived from a continuous utterance of "Let's go shop (chop)." This finding is in apparent contradiction to that of Dorman et al. (1979) described in the preceding paragraph. Dechovitz et al. interpreted it as showing that dynamic information for utterance continuity may override a perceived change in source (despite the concomitant auditory discontinuities). If this interpretation is correct, it may point to another instance where purely auditory principles fail to explain phonetic perception. Some of the variables that determine the perceived continuity of an utterance are likely to be auditory (cf. Bregman, 1978); however, there may also be speech-specific factors that reflect what listeners consider plausible and possible in the dynamic context of natural utterances.

CONCLUSIONS

The findings reviewed above provide a wealth of results that, in large measure, cannot be accounted for by our current knowledge of auditory psychophysics. Although there remains much to be learned about the perception of complex auditory stimuli, some trading relations and context effects seem *a priori* unlikely to reflect an auditory level of interaction, and at least one--audio-visual integration--simply cannot derive from that level. While efforts to delineate the role of general auditory processes in speech perception should certainly continue, it may be predicted that this role will be restricted largely to the perception of nonphonetic stimulus attributes.

This is not to say that auditory properties of the signal are not the basic carrier of the linguistic message. However, auditory psychophysics gains knowledge about the perception of these properties in large part from listeners' judgments in psychophysical experiments, and these judgments are made in a different frame of reference from the judgments of speech. Auditory variables, but *not* auditory judgments, are the basis of phonetic perception. Even those limitations imposed by the auditory system that have to do with detectability and resolution may not play any important role in phonetic distinctions. For instance, there is no reason why phonetic category boundaries could not be placed at suprathreshold auditory parameter settings that seem arbitrary from a psychophysical viewpoint but are well motivated by the articulatory and acoustic patterns that characterize a given language. And even though phonetic and auditory boundaries may sometimes coincide, there is the more fundamental question whether such "boundaries" play any role in the perception of natural speech, considering the fact that natural speech is different in a number of ways from the artificial stimuli employed in speech discrimination tasks. While the objection of ecologically invalid stimuli extends to most of the studies reviewed in this paper, the present emphasis has been on *processes of perceptual integration* that promise to be more general than static concepts such as boundary locations.

Two possible criticisms of the research reviewed here should be mentioned. One is that nearly all studies demonstrated perceptual integration in situations of high uncertainty produced by ambiguous settings of the primary cue(s) for a given phonetic distinction. The perceptual integration observed may have been motivated by that ambiguity. In that case, it may be that perceptual integration does not occur to the same extent in natural situations, where the primary cues are often sufficient for accurate phonetic perception.

The other criticism is that, although the trading relations and context effects reviewed here have been described as complex interactions between separate cues, it may well be that these cues do not function as perceptual entities that are "extracted" and then recombined into a unitary phonetic percept (cf. Bailey & Summerfield, 1980). In that view, cues serve only descriptive purposes; the perceptual interactions between them can be understood as resulting from the listeners' apprehension of the articulatory events they convey. While cues (i.e., acoustic segments) are indispensable for describing how the articulatory information is represented in the signal, we need not postulate special perceptual processes that construct or derive the articulatory information from these elementary pieces. Rather, the articulatory information may be said to be directly available (Gibson, 1966; Neisser, 1976). This is an attractive proposal; however, we should not forget that there are real questions to be answered about the mechanisms that accomplish phonetic perception and that we know so woefully little about at present. If cues and their interactions have no place in a description of these mechanisms, we are faced with the more fundamental problem of finding the proper ingredients for a model of speech perception.

There is reason to believe that the information processing approaches currently in vogue are not likely to lead us very far in that regard. To understand how our perceptual systems work, we need to understand how a complex biological system (our brain) integrates and differentiates information, how it is modified by experience, and how the structure of the input (i.e., the environment) gets to be represented in the system. These are complex biological questions whose solution will not come easily. Computer analogies are largely tautological and distract from the fundamental biological and philosophical problems that lie at the heart of the problem of perception (see, e.g., Hayek, 1952; Piaget, 1967; Studdert-Kennedy, in press, Note 9). In a particularly enlightening discussion, Fodor (Note 10) has recently argued for the modularity of the speech (and language) system, i.e., for its specificity and relative isolation from other perceptual and cognitive systems. He also pointed out that it is precisely such modular systems that we have some hope of understanding, whereas explanations of perception in terms of general principles remains interminably *ad hoc*. Thus, we should not be surprised to find that speech perception is accomplished by means entirely particular to that mode. The problem of how to investigate and describe those means will keep us busy for some time to come.

REFERENCE NOTES

1. Grunke, M. E., & Pisoni, D. B. Some experiments on perceptual learning of mirror-image acoustic patterns. *Research on Speech Perception* (Department of Psychology, Indiana University), 1979, *5*, 147-182.

2. Serniclaes, W. La simultanéité des indices dans la perception du voisement des occlusives. *Rapport d'Activités de l'Institut de Phonétique* (Bruxelles: Université Libre), 1973, *7*(2), 59-67.
3. Serniclaes, W. Traitement indépendant ou interaction dans le processus de structuration perceptive des indices de voisement? *Rapport d'Activités de l'Institut de Phonétique* (Bruxelles: Universite Libre), 1975, *9*(2), 47-57.
4. Miller, J. L., & Eimas, P. D. *Contextual perception of voicing by infants*. Paper presented at the Biennial Meeting of the Society for Research in Child Development, Boston, MA, April 1981.
5. Kunisaki, O., & Fujisaki, H. On the influence of context upon perception of voiceless fricative consonants. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics* (University of Tokyo), 1977, *11*, 85-91.
6. Zue, V. W. *Acoustic characteristics of stop consonants: A controlled study*. Technical Report No. 523, Lincoln Laboratory (Massachusetts Institute of Technology, Lexington, Massachusetts), 1976.
7. Carrell, T. D., Pisoni, D B., & Gans, S. J. Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech. *Research on Speech Perception* (Department of Psychology, Indiana University), 1980, *6*, 421-436.
8. Shigeno, S., & Fujisaki, H. Context effects in phonetic and non-phonetic vowel judgments. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics* (University of Tokyo), 1980, *14*, 217-224.
9. Studdert-Kennedy, M. *Are utterances prepared and perceived in parts? Perhaps*. Paper presented at the First International Conference on Event Perception, University of Connecticut, Storrs, June 1981.
10. Fodor, J. A. *The modularity of mind*. Unpublished paper, M.I.T., 1981.

REFERENCES

Aslin, R. N., & Pisoni, D. B. Some developmental processes in speech perception. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology. Vol. 2: Perception.* New York: Academic Press, 1980. Pp. 67-96.

Bailey, P. J., & Summerfield, Q. Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, *6*, 536-563.

Bailey, P. J., Summerfield, Q., & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research*, 1977, *SR-51/52*, 1-25.

Barton, D. Phonemic perception in children. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology. Vol. 2: Perception.* New York: Academic Press, 1980. Pp. 97-116.

Best, C. T., Morrongiello, B., & Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, 1981, *29*, 191-211.

Blechner, M. J. *Musical skill and the categorical perception of harmonic mode*. Unpublished doctoral dissertation, Yale University, 1977.

Blechner, M. J., Day, R. S., & Cutting, J. E. Processing two dimensions of nonspeech stimuli: The auditory-phonetic distinction reconsidered. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, *2*, 257-266.

Blumstein, S. E., & Stevens, K. N. Acoustic invariance in speech production. *Journal of the Acoustical Society of America*, 1979, *66*, 1001-1017.

Blumstein, S. E., & Stevens, K. N. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America*, 1980, *67*, 648-662.

Bradshaw, J. L., & Nettleton, N. C. The nature of hemispheric specialization in man. *The Behavioral and Brain Sciences*, 1981, *4*, 51-63.

Bregman, A. S. The formation of auditory streams. In J. Requin (Ed.), *Attention and performance VII.* Hillsdale, NJ: Erlbaum, 1978. Pp. 63-75.

Burns, E. M., & Ward, W. D. Categorical perception -- phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America*, 1978, *63*, 456-468.

Carden, G., Levitt, A. G., Jusczyk, P. W., & Walley, A. Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics*, 1981, *29*, 26-36.

Cole, R. A., & Scott, B. Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, 1973, *27*, 441-449.

Cutting, J. E. Two left-hemisphere mechanisms in speech perception. *Perception & Psychophysics*, 1974, *16*, 601-612.

Cutting, J. E. Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, 1976, *83*, 114-140.

Cutting, J. E. There may be nothing peculiar to perceiving in a speech mode. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Erlbaum, 1978. Pp. 229-244.

Cutting, J. E., & Rosner, B. S. Categories and boundaries in speech and music. *Perception & Psychophysics*, 1974, *16*, 564-570.

Darwin, C. J., & Bethell-Fox, C. E. Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 665-672.

Day, R. S. *Fusion in dichotic listening*. Unpublished doctoral dissertation, Stanford University, 1968.

Dechovitz, D. R., Rakerd, B., & Verbrugge, R. R. Effects of utterance continuity on phonetic judgments. *Haskins Laboratories Status Report on Speech Research*, 1980, *SR-62*, 101-116.

Denes, P. Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 1955, *27*, 761-764.

Derr, M. A., & Massaro, D. W. The contribution of vowel duration, F_o contour, and frication duration as cues to the /juz/-/jus/ distinction. *Perception & Psychophysics*, 1980, *27*, 51-59.

Diehl, R. L. Feature detectors for speech: A critical reappraisal. *Psychological Bulletin*, 1981, *89*, 1-18.

Diehl, R. L., Souther, A. F., & Convis, C. L. Conditions on rate normalization in speech perception. *Perception & Psychophysics*, 1980, *27*, 435-443.

Divenyi, P. L. Some psychoacoustic factors in phonetic analysis. *Proceedings of the Ninth International Congress of Phonetic Sciences*, Vol. II. Copenhagen: University of Copenhagen, 1979. Pp. 445-452.

Dorman, M. F., & Raphael, L. J. Distribution of acoustic cues for stop consonant place of articulation in VCV syllables. *Journal of the Acoustical Society of America*, 1980, *67*, 1333-1335.

Dorman, M. F., Raphael, L. J., & Isenberg, D. Acoustic cues for a fricative-affricate contrast in word-final position. Journal of Phonetics, 1980, 8, 397-405.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. Journal of the Acoustical Society of America, 1979, 65, 1518-1532.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception & Psychophysics, 1977, 22, 109-122.

Eimas, P. D., & Corbit, J. D. Selective adaptation of linguistic feature detectors. Cognitive Psychology, 1973, 4, 99-109.

Eimas, P. D., & Miller, J. L. Contextual effects in speech perception. Science, 1980, 209, 1140-1141.

Fitch, H. L. Distinguishing temporal information for speaking rate from temporal information for intervocalic stop consonant voicing. Haskins Laboratories Status Report on Speech Research, 1981, SR-65, 1-32.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. Perception & Psychophysics, 1980, 27, 343-350.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. The identification and discrimination of synthetic vowels. Language and Speech, 1962, 5, 171-189.

Fujisaki, H., & Kunisaki, O. Analysis, recognition, and perception of voiceless fricative consonants in Japanese. IEEE Transactions (ASSP), 1978, 26, 21-27.

Fujisaki, H., & Shigeno, S. Context effects in the categorization of speech and nonspeech stimuli. In J. J. Wolf & D. H. Klatt (Eds.), Speech communication papers presented at the 97th Meeting of the Acoustical Society of America. New York: Acoustical Society of America, 1979.

Ganong, W. F. III The selective adaptation effects of burst-cued stops. Perception & Psychophysics, 1978, 24, 71-83.

Gerstman, L. Cues for distinguishing among fricatives, affricates, and stop consonants. Unpublished doctoral dissertation, New York University, 1957.

Gibson, J. J. The senses considered as perceptual systems. Boston, Mass.: Houghton-Mifflin, 1966.

Grosjean, F., & Lane, H. How the listener integrates the components of speaking rate. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 538-543.

Haggard, M. P., Ambler, S., & Callow, M. Pitch as a voicing cue. Journal of the Acoustical Society of America, 1970, 47, 613-617.

Halle, M., & Stevens, K. N. Analysis by synthesis. In W. Wathen-Dunn & L. E. Woods (Eds.), Proceedings of the seminar on speech compression and processing, Vol. 2. AFCRC-TR-59-198, USAF Cambridge Research Center, 1959.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., & Cooper, F. S. Effect of third-formant transitions on the perception of the voiced stop consonants. Journal 3 of the Acoustical Society of America, 1958, 30, 122-126.

Hasegawa, A. Some perceptual consequences of fricative coarticulation. Unpublished doctoral dissertation, Purdue University, 1976.

Hayek, F. A. The sensory order. Chicago: University of Chicago Press, 1952.

Healy, A. F., & Repp, B. H. Context sensitivity and phonetic mediation in categorical perception. Journal of Experimental Psychology: Human Perception and Performance, in press.

Heuven, V. J. van The relative contribution of rise time, steady time, and overall duration of noise bursts to the affricate-fricative distinction in English: A re-analysis of old data. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers presented at the 97th Meeting of the Acoustical Society of America*. New York: Acoustical Society of America, 1979.

Hoffman, H. S. Study of some cues in the perception of the voiced stop consonants. *Journal of the Acoustical Society of America*, 1958, *30*, 1035-1041.

Hogan, J. T., & Rozsypal, A. J. Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America*, 1980, *67*, 1764-1771.

House, A. S., Stevens, K. N., Sandel, T. T., & Arnold, J. B. On the learning of speechlike vocabularies. *Journal of Verbal Learning and Verbal Behavior*, 1962, *1*, 133-143.

Isenberg, D., & Liberman, A. M. Speech and nonspeech percepts from the same sound. *Journal of the Acoustical Society of America*, 1978, *64* (Supplement No. 1), S20. (Abstract)

Jusczyk, P. W. Infant speech perception: A critical appraisal. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, NJ: Erlbaum, 1981. Pp. 113-164.

Kewley-Port, D. *Representations of spectral change as cues to place of articulation of stop consonants*. Unpublished doctoral dissertation, City University of New York, 1981.

Kohler, K. J. Dimensions in the perception of fortis and lenis plosives. *Phonetica*, 1979, *36*, 332-343.

Klatt, D. H. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 1976, *59*, 1208-1221.

Kuhl, P. K. Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories. *Journal of the Acoustical Society of America*, 1981, *70*, 340-349.

Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, 1978, *63*, 905-917.

Ladefoged, P., & Broadbent, D. E. Information conveyed by vowels. *Journal of the Acoustical Society of America*, 1957, *29*, 98-104.

Liberman, A. M. Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. *Proceedings of the Ninth International Congress of Phonetic Sciences*, Vol. II. Copenhagen: University of Copenhagen, 1979. Pp. 468-473.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497-516.

Liberman, A. M., Harris, K. S., Eimas, P. D., Lisker, L., & Bastian, J. An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 1961, *4*, 175-195.

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. The discrimination of relative onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 1961, *61*, 379-388.

Liberman, A. M., Isenberg, D., & Rakerd, B. Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, 1981, *30*, 133-143.

Lindner, G. Veraenderung der Beurteilung synthetischer Vokale unter dem Einfluss des Sukzessivkontrastes. *Zeitschrift fuer Phonetik, Sprachwissenschaft, und Kommunikationsforschung*, 1966, *19*, 287-307.

Lisker, L. Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 1957, *33*, 42-49.

Lisker, L. Is it VOT or a first-formant transition detector? *Journal of the Acoustical Society of America*, 1975, *57*, 1547-1551.

Lisker, L. Closure hiatus: Cue to voicing, manner and place of consonant occlusion. *Haskins Laboratories Status Report on Speech Research*, 1978, *SR-53*, (Vol. 1), 79-86. (a)

Lisker, L. *Rapid* vs. *rabid*: A catalogue of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Research*, 1978, *SR-54*, 127-132. (b)

Lisker, L. On buzzing the English /b/. *Haskins Laboratories Status Report on Speech Research*, 1978, *SR-55/56*, 181-188. (c)

Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., & Mandler, R. On pushing the voice-onset-time (VOT) boundary about. *Language and Speech*, 1977, *20*, 209-216.

Lisker, L., & Price, P. J. Context-determined effects of varying closure duration. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers presented at the 97th Meeting of the Acoustical Society of America*. New York: Acoustical Society of America, 1979.

MacDonald, J., & McGurk, H. Visual influences on speech perception processes. *Perception & Psychophysics*, 1978, *24*, 253-257.

Mann, V. A. Influence of preceding liquid on stop consonant perception. *Perception & Psychophysics*, 1980, *28*, 407-412.

Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. Further investigation into the influence of preceding liquids on stop consonant perception. *Journal of the Acoustical Society of America*, 1981, *69* (Supplement No. 1), S91. (Abstract)

Mann, V. A., & Repp, B. H. Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, 1980, *28*, 213-228.

Mann, V. A., & Repp, B. H. Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, 1981, *69*, 548-558.

Massaro, D. W., & Cohen, M. M. The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinctions. *Journal of the Acoustical Society of America*, 1976, *60*, 704-717.

Massaro, D. W., & Cohen, M. M. Voice onset time and fundamental frequency as cues to the /zi/ - /si/ distinction. *Perception & Psychophysics*, 1977, *22*, 373-382.

Mattingly, I. G., & Levitt, A. G. Perception of stop consonants before low unrounded vowels. *Haskins Laboratories Status Report on Speech Research*, 1980, *SR-61*, 167-174.

Mattingly, I. G., Liberman, A. M., Syrdal, A. M., & Halwes, T. Discrimination in speech and nonspeech modes. *Cognitive Psychology*, 1971, *2*, 131-157.

May, J. Vocal tract normalization for /s/ and /š/. *Haskins Laboratories Status Report on Speech Research*, 1976, *SR-48*, 67-73.

McGurk, H., & McDonald, J. Hearing lips and seeing voices. *Nature*, 1976, *264*, 746-748.

Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J., & Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 60, 410-417.

Miller, J. L. Contextual effects in the discrimination of stop consonant and semivowel. Perception & Psychophysics, 1980, 28, 93-95.

Miller, J. L. The effect of speaking rate on segmental distinctions: Acoustic variation and perceptual compensation. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

Miller, J. L., & Grosjean, F. How the components of speaking rate influence perception of phonetic segments. Journal of Experimental Psychology: Human Perception and Performance, 1981, 7, 208-215.

Miller, J. L., & Liberman, A. M. Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception & Psychophysics, 1979, 25, 457-465.

Morse, P. A. The infancy of infant speech percception: The first decade of reseach. Brain, Behavior, and Evolution, 1979, 16, 351-373.

Neisser, U. Cognition and reality. San Francisco: Freeman, 1976.

Pastore, R. E. Contralateral cueing effects in the perception of aspirated stop consonants. Journal of the Acoustical Society of America, 1978, 64 (Supplement No. 1), S17. (Abstract)

Pastore, R. E. Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, N.J.: Erlbaum, 1981.

Pastore, R. E., Ahroon, W. A., Baffuto, K. J., Friedman, C., Puleo, J. S., & Fink, E. A. Common factor model of categorical perception. Journal of Experimental Psychology: Human Perception and Performance, 1977, 3, 686-696.

Pastore, R. E., Ahroon, W. A., Puleo, J. S., Crimmins, D. B., Golowner, D. B., & Berger, R. S. Processing interactions between two dimensions of nonphonetic auditory signals. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2, 267-276.

Pastore, R. E., Harris, L. B., & Kaplan, J. TOT: An acoustic CV syllable temporal onset analog? Journal of the Acoustical Society of America, 1981, 69 (Supplement No. 1), S93. (Abstract)

Piaget, J. Biology and knowledge. Chicago: University of Chicago Press, 1967.

Pickett, J. M., & Decker, L. R. Time factors in perception of a double consonant. Language and Speech, 1960, 3, 11-17.

Pisoni, D. B. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Pisoni, D. B. Adaptation of the relative onset time of two-component tones. Perception & Psychophysics, 1980, 28, 337-346.

Pols, L. C. W., & Schouten, M. E. H. Identification of deleted consonants. Journal of the Acoustical Society of America, 1978, 64, 1333-1337.

Price, P. J., & Lisker, L. (/b/→/p/) but ~(/p/→/b/). In J. J. Wolf & D. H. Klatt (Eds.), Speech communication papers presented at the 97th Meeting of the Acoustical Society of America. New York: Acoustical Society of America, 1979.

Rand, T. C. Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research, 1971, SR-25/26, 141-146.

Rand, T. C. Dichotic release from masking for speech. Journal of the Acoustical Society of America, 1974, 55, 678-680.

Raphael, L. J. Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. Journal of the Acoustical Society of America, 1972, 51, 1296-1303.

Raphael, L. J. Durations and contexts as cues to word-final cognate opposition in English. Phonetica, 1981, 38, 126-147.

Remez, R. E. Adaptation of the category boundary between speech and nonspeech: A case against feature detectors. Cognitive Psychology, 1979, 11, 38-57.

Remez, R. E., Cutting, J. E., & Studdert-Kennedy, M. Cross-series adaptation using song and string. Perception & Psychophysics, 1980, 27, 524-530.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.

Repp, B. H. Dichotic "masking" of voice onset time. Journal of the Acoustical Society of America, 1976, 59, 183-194.

Repp, B. H. Perceptual integration and differentiation of spectral cues for intervocalic stop consonants. Perception & Psychophysics, 1978, 24, 471-485. (a)

Repp, B. H. Interdependence of voicing and place decisions for stop consonants in initial position. Haskins Laboratories Status Report on Speech Research, 1978, SR-53 (Vol. II), 117-150. (b)

Repp, B. H. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech, 1979, 22, 173-189.

Repp, B. H. Bidirectional contrast effects in the perception of VC-CV sequences. Haskins Laboratories Status Report on Speech Research, 1980, SR-63/64, 157-176. (a)

Repp, B. H. Perception and production of two-stop-consonant sequences. Haskins Laboratories Status Report on Speech Research, 1980, SR-63/64, 177-194. (b)

Repp, B. H. Two strategies in fricative discrimination. Perception & Psychophysics, 1981, 30, 217-227. (a)

Repp, B. H. Auditory and phonetic trading relations between acoustic cues in speech perception: Preliminary results. Haskins Laboratories Status Report on Speech Research, 1981, SR-67/68, this volume. (b)

Repp, B. H., Healy, A. F., & Crowder, R. G. Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance, 1979, 5, 129-145.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.

Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulation. Journal of the Acoustical Society of America, 1981, 69, 1154-1163. (a)

Repp, B. H., & Mann, V. A. Fricative-stop coarticulation: Acoustic and perceptual evidence. Haskins Laboratories Status Report on Speech Research, 1981, SR-67/68, this volume. (b)

Roberts, M., & Summerfield, Q. Audio-visual adaptation in speech perception. Perception & Psychophysics, 1981, 30, 309-314.

Rosen, S., & Howell, P. Plucks and bows are not categorically perceived. Perception & Psychophysics, 1981, 30, 156-168.

Sawusch, J. R., & Jusczyk, P. Adaptation and contrast in the perception of voicing. Journal of Experimental Psychology: Human Perception and Performance, 1981, 7, 408-421.

Sawusch, J. R., & Pisoni, D. B. Category boundaries for speech and nonspeech sounds. *Journal of the Acoustical Society of America*, 1973, *54*, 76. (Abstract)

Sawusch, J. R., Pisoni, D. B., & Cutting, J. E. Category boundaries for linguistic and non-linguistic dimensions of the same stimuli. *Journal of the Acoustical Society of America*, 1974, *55* (Supplement No. 1), S55. (Abstract)

Schouten, M. E. H. The case against a speech mode of perception. *Acta Psychologica*, 1980, *44*, 71-98.

Schwartz, M. F. Identification of speaker sex from isolated voiceless fricatives. *Journal of the Acoustical Society of America*, 1968, *43*, 1178-1179.

Searle, C. L., Jacobson, J. Z., & Rayment, S. G. Stop consonant discrimination based on human audition. *Journal of the Acoustical Society of America*, 1979, *65*, 799-809.

Serniclaes, W. Perceptual processing of acoustic correlates of the voicing feature. *Proceedings of the Speech Communication Seminar*. Stockholm, 1974. Pp. 87-93.

Siegel, J. A., & Siegel, W. Categorical perception of tonal intervals: Musicians can't tell *sharp* from *flat*. *Perception & Psychophysics*, 1977, *21*, 399-407.

Simon, C., & Fourcin, A. J. Cross-language study of speech-pattern learning. *Journal of the Acoustical Society of America*, 1978, *63*, 925-935.

Smith, M. R. Perception of word stress and syllable length. *Journal of the Acoustical Society of America*, 1978, *63* (Supplement No. 1), S55. (Abstract)

Soli, S. D. Structure and duration of vowels together specify fricative voicing. *Journal of the Acoustical Society of America*, in press.

Stevens, K. N. The potential role of property detectors in the perception of consonants. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech*. New York: Academic Press, 1975, 303-330.

Stevens, K. N., & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 1978, *64*, 1358-1368.

Stevens, K. N., & Klatt, D. H. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 1974, *55*, 653-659.

Strange, W., Verbrugge, R., Shankweiler, D. P., & Edman, T. R. Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 1976, *60*, 213-224.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics*. New York: Academic Press, 1976, 243-293.

Studdert-Kennedy, M. A note on the biology of speech perception. In J. Mehler, M. Garrett, & E. Walker (Eds.), *Perspectives in mental representation*. Hillsdale, N.J.: Erlbaum, in press.

Studdert-Kennedy, M. Cerebral hemispheres: Specialized for the analysis of what? *The Behavioral and Brain Sciences*, 1981, *4*, 76-77.

Summerfield, A. Q. *Information processing analyses of perceptual adjustments to source and context variables in speech*. Unpublished doctoral dissertation, Queen's University of Belfast, 1975.

Summerfield, Q. Use of visual information for phonetic perception. *Phonetica*, 1979, *36*, 314-331.

Summerfield, Q. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, *7*, 1074-1095.

Summerfield, Q. Does VOT equal TOT or NOT? Examination of a possible auditory basis for the perception of voicing in initial stops. *Journal of the Acoustical Society of America*, in press.

Summerfield, Q., Bailey, P. J., Seton, J., & Dorman, M. F. Fricative envelope parameters and silent intervals in distinguishing 'slit' and 'split.' *Phonetica*, 1981, *38*, 181-192.

Summerfield, A. Q., & Haggard, M. P. Perceptual processing of multiple cues and contexts: Effects of following vowel upon stop consonant voicing. *Journal of Phonetics*, 1974, *2*, 279-295.

Summerfield, A. Q., & Haggard, M. P. Vocal tract normalization as demonstrated by reaction times. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech*. London: Academic Press, 1975, 115-142.

Summerfield, Q., & Haggard, M. P. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, 1977, *62*, 435-448.

Thompson, C. L., & Hollien, H. Some contextual effects on the perception of synthetic vowels. *Language and Speech*, 1970, *13*, 1-13.

Walley, A. C., Pisoni, D. B., & Aslin, R. N. The role of early experience in the development of speech perception. In R. N. Aslin, J. Alberts, & M. R. Petersen (Eds.), *Sensory and perceptual development*. New York: Academic Press, 1981.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. Auditory sequence: Confusion of patterns other than speech or music. *Science*, 1969, *164*, 586-587.

Waters, R. S., & Wilson, W. A., Jr. Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. *Perception & Psychophysics*, 1976, *19*, 285-289.

Whalen, D. H. Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America* 1981, *69*, 275-282.

Wolf, C. G. Voicing cues in English final stops. *Journal of Phonetics*, 1978, *6*, 299-309.

Wood, C. C. Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *104*, 3-20.

Zatorre, R. J., & Halpern, A. R. Identification, discrimination, and selective adaptation of simultaneous musical intervals. *Perception & Psychophysics*, 1979, *26*, 384-395.

Zlatin, M. A. Voicing contrast: Perceptual and productive voice onset time characteristics of adults. *Journal of the Acoustical Society of America*, 1974, *56*, 981-994.

Zwicker, E., Terhardt, E., & Paulus, E. Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America*, 1979, *65*, 487-498.

FOOTNOTES

[1] A rule of thumb for distinguishing a trading relation from a context effect is that the phonetic equivalence resulting from a trading relation is strong in the sense that two phonetically equivalent stimuli (syllables or words) are difficult to tell apart (Fitch et al., 1980), whereas the phonetic equivalence produced by trading a critical cue against some contextual influence is restricted to the target segment, as it always involves a readily detectable change in one or more contextual segments. To the extent that a change in context (e.g., vowel quality) also modifies critical cues (e.g., formant transitions), context effects may sometimes include disguised trading relations.

[2] The attempt to define integrated cues must be distinguished from independent efforts to represent the speech signal in a way that takes into account peripheral auditory transformations (Searle, Jacobson, & Rayment, 1979; Zwicker, Terhardt, & Paulus, 1979). Such representations are, of course, very useful and may lead to the redefinition of some cues; however, they do not, by themselves, solve the problem of cue definition.

[3] In essence, this kind of study investigates whether multidimensionally varying speech stimuli are perceived categorically. Traditional studies of categorical perception have been exclusively concerned with stimuli varying on a single dimension, or varying on several dimensions in a perfectly correlated fashion. Note that, in these studies, physically different stimuli from the region of the category boundary are not phonetically equivalent--they have different response distributions. As soon as two or more cues are varied, however, pairs of phonetically equivalent stimuli can be found for any given response distribution. Thus, the influence of phonetic categorization on discrimination judgments can be factored out, at least in principle (see Footnote 4).

[4] To produce precise (rather than just average) phonetic equivalence, it would not only be necessary to take into account the fact that individual listeners show trading relations of varying magnitude but also that (covert) labeling responses may change in the context of a discrimination task (Repp et al., 1979). Thus, the stimulus parameters would have to be adjusted separately for each listener, based on labeling data collected with the stimulus sequences of the discrimination task. This procedure would optimize the opportunity to verify the prediction that stimuli in the conflicting-cues condition are more difficult to discriminate than those in the cooperating-cues condition, with the single-cue condition in between. However, this order of difficulty is likely to obtain also when the choices of parameters are less than optimal.

[5] Most interestingly, the only completed study (so far) of a trading relation in human infants (Miller & Eimas, Note 4) has yielded a positive result: The boundary on a VOT continuum was significantly affected by the duration of the formant transitions, a variable that is confounded with F1 onset frequency (cf. Summerfield & Haggard, 1977). Kuhl & Miller (1978) obtained a similar result with chinchillas. This trading relation, at least, appears to be of auditory origin, even though the principle involved is not yet clear. It seems likely, though, that not all trading relations will follow this pattern.

[6] That the subjects focused on one cue only was a strategy furthered by the AXB classification task of Best et al. In a different paradigm, the subjects may pay attention to both cues at the same time (cf. Repp, 1981b). The important point is that, in the auditory mode, the cues are not integrated into a unitary percept, so that listeners may choose between selective-attention and divided-attention strategies.

[7] In that connection, the study of Simon and Fourcin (1978) might be mentioned, which showed that the trading relation between VOT and F1 transition trajectory as cues to stop consonant voicing emerged at age 4 in British children but was absent in 2- and 3-year olds. Recently, however, Miller & Eimas (Note 4) found a related trading relation (between VOT and transition duration) in American infants. This conflict needs to be resolved.

[8] I have experienced this effect myself (together with a number of my colleagues at Haskins) and can confirm that it is a true perceptual phenomenon, not some kind of inference or bias in the face of conflicting information. The observer really believes that he or she hears what, in fact, he or she only sees on the screen; there is little or no awareness of anything odd happening. However, the effect is not always that strong; its presence and strength depend on the particular combination of syllables, in a way that can also, in part, be explained by reference to articulation. It is strongest when the visual information makes the auditory information impossible in articulatory terms. The details of the effect and of the relevant variables remain to be investigated.

[9] These experiments concern the disruption of perceptual integration of cues. However, context effects can presumably be similarly blocked by a change in apparent source. Diehl, Souther, and Convis (1980) recently reported a study in which a rate normalization effect (of a precursor on the /ga/-/ka/ distinction) was eliminated by a change of voice. Unfortunately, their data were not entirely consistent and call for replication.

TEMPORAL PATTERNS OF COARTICULATION: LIP ROUNDING*

Fredericka Bell-Berti+ and Katherine S. Harris++

Abstract. According to some theories, anticipatory coarticulation occurs when phones for which a feature is unspecified precede one for which the feature is specified, with consequent migration of the feature value to the antecedent phones. Carryover coarticulation, on the other hand, is often attributed to "articulatory sluggishness." In this paper, EMG evidence is provided that this formulation is inadequate, since the beginning of EMG activity associated with vowel lip rounding is independent of measures of the acoustic duration of adjacent consonants. We suggest that the often noted vowel-rounding gesture simply co-occurs during predictable intervals with portions of preceding and following lingual consonant articulations.

INTRODUCTION

A central problem in understanding the relationship between speech production and perception is the disparity between the perceptual representation of speech as a series of discrete events, composed of partially commutable elements, and the acoustic representation as a continuously varying stream, without obvious phonetic segment markers. This acoustic stream is generated by the activity of the several articulators, whose activity is apparently continuous and context dependent. Many theories of coarticulation attempt to solve the problem of context sensitivity by positing some kind of speech synthesis process that occurs in production, and allows the fitting together of the discrete units into the continuous stream. The task of the theorist, then, is to write the adjustment rules.

In a widely cited theory of anticipatory coarticulation, Henke (1966) provides a fairly typical formulation. Each phone in an articulatory string

*In press, Journal of the Acoustical Society of America. Some of these data were presented at the 96th Meeting of the Acoustical Society of America, Honolulu, Hawaii, November 1978 [Journal of the Acoustical Society of America, 64, S92(A), 1978], and at the 97th Meeting of the Acoustical Society of America, Cambridge, Massachusetts, June 1979 [Journal of the Acoustical Society of America, 65, S22(A), 1979].

+Also St. John's University, Jamaica, New York 11439.

++Also The Graduate School, The City University of New York, New York, New York 10036.

Acknowledgment. This work was supported by NINCDS grants NS-13617 and NS-05332 and BRS grant RR-05596 to the Haskins Laboratories.

is conceived as composed of a bundle of articulatory features. Anticipatory coarticulation occurs when phones for which a given feature is unspecified precede one for which the feature is specified, with consequent subjection of the antecedent phones to the feature value of the following phone. Since time is unspecified in the theory, the temporal duration occupied by the string of antecedent phones is presumably irrelevant; all will acquire the same feature value.

It has been claimed by Fowler (1980) that all such theories of coarticulation belong to the class of extrinsic timing models of speech production. Such models assume that the dimension of time is excluded from the specification of a phonological segment in the motor plan for the utterance. In Fowler's view, such accounts must therefore necessarily fail to explain or predict coarticulation. While one may or may not accept her argument in its larger theoretical framework, we believe that purely substantive evidence can be marshaled against such phonological segment theories as a class.

In an earlier report (Bell-Berti & Harris, 1979) we provided evidence that this formulation is inadequate, and have elsewhere suggested an alternative hypothesis (Bell-Berti, 1980; Bell-Berti & Harris, 1981). Specifically, we found that if a rounded vowel was preceded by one or two consonants presumably unspecified for rounding, the electromyographic activity associated with rounding began a constant time, rather than a constant number of segments, before the onset of the vowel.

The present experiment was designed to extend the earlier one in several ways. First, we have examined both anticipatory and carryover coarticulation of lip rounding. Often, "articulatory sluggishness" explanations are proposed for carryover coarticulation while "planning" explanations are proposed for anticipatory coarticulation (e.g., MacNeilage, 1970). However, if both anticipatory and carryover effects appear to be guided by the same articulatory rules, disparate explanations for these two effects seem less plausible.

Secondly, we have examined the special case in which coarticulation occurs from one vowel to another vowel, where both vowels are rounded and are separated by intervening consonants without rounding specification. In such cases, it has been shown that a "trough" will occur--that is, EMG activity will be reduced at some point in the vowel-to-vowel period. This situation is, of course, not explicable by the type of model of coarticulation exemplified by Henke's, as we (Bell-Berti & Harris, 1974) and others (e.g., Gay, 1978) have pointed out.

Thirdly, we extended the design of the experiment to include longer strings of consonants preceding or following the rounded vowel than the original maximum of two-element clusters. We also increased the subject pool, and included subjects naive to the purposes of the experiment.

Fourthly, we checked the subjects to see if orbicularis oris activity occurred for segment sequences for which no lip rounding was specified. In a theory like Henke's, it is assumed that a feature, such as lip rounding, spreads from a phone for which it is specified, to the preceding phones for which it is not. If the preceding phones carry a specification for the feature, the experiment provides no test of the theory. Earlier studies (Daniloff & Moll, 1968) have been criticized by later authors (Benguerel & Cowan, 1974) for possible design flaws of this type. For the experiment described here, we assume that the alveolars, especially /s/, are neutral with respect to rounding. Hence, we would expect that in sequences of the form /isi/, no EMG evidence of rounding would be observed, since the vowel /i/ is traditionally characterized as spread, and the consonant /s/ is not traditionally characterized with respect to lip rounding (Bronstein, 1960). However, since traditional descriptions are often incomplete concerning fine-grained articulatory detail, it seemed worthwhile to make an explicit check of lip activity during the sequence /isi/ for each speaker.

As in the previous study, we have used an electromyographic indicator of rounding, the activity of the orbicularis oris muscle. The relationship between orbicularis oris activity and vowel rounding is well documented by a number of studies (Harris, Lysaught, & Schvey, 1965; Fromkin, 1966; Tatham & Morton, 1968; Sussman & Westbury, 1981).

METHODS

Speech Materials

The experimental speech materials were two-word phrases spoken within the carrier phrase "It's a __________ again." The first word was one from the set "lee, lease, leased, loo, loose, loosed," while the second word was one from the set "tool, stool, teal, steel." All utterances whose second word was either "tool" or "stool" will be called the "anticipatory" set in the discussion below, since they were designed to examine anticipatory lip rounding. Conversely, those utterances whose first word was "loo," "loose," or "loosed" and whose second word was "teal" or "steel" will be called the "carryover" set.

In additon to these eighteen experimental utterances (12 in the anticipatory and six in the carryover sets), we examined an additional group that included "lee teal" and "lee seal," to determine whether a speaker produced either or both of the alveolar consonants /t/ or /s/ with orbicularis oris EMG activity, in the absence of a rounded vowel.

The experimental utterances were placed in randomized lists that included additional items intended as foils. Five subjects read the randomized lists until 14 to 18 repetitions of each experimental utterance had been recorded. A sixth subject produced only ten repetitions of each utterance type. Subjects were asked to read the sentences from an orthographic representation, and, thus, produced the phonetic sequences natural to the word combinations -- e.g., [listul] rather than [list tul] for "leased tool."

Subjects

Five of the six subjects of the experiment were naive to its purposes; the sixth subject, the senior author, had also been a subject in the previous study. Three of the five naive subjects showed activity in orbicularis oris associated with the production of /s/ in spread vowel environments; hence, their data were not further analyzed. All of the subjects whose data are presented here are speakers of educated Greater Metropolitan New York City English.

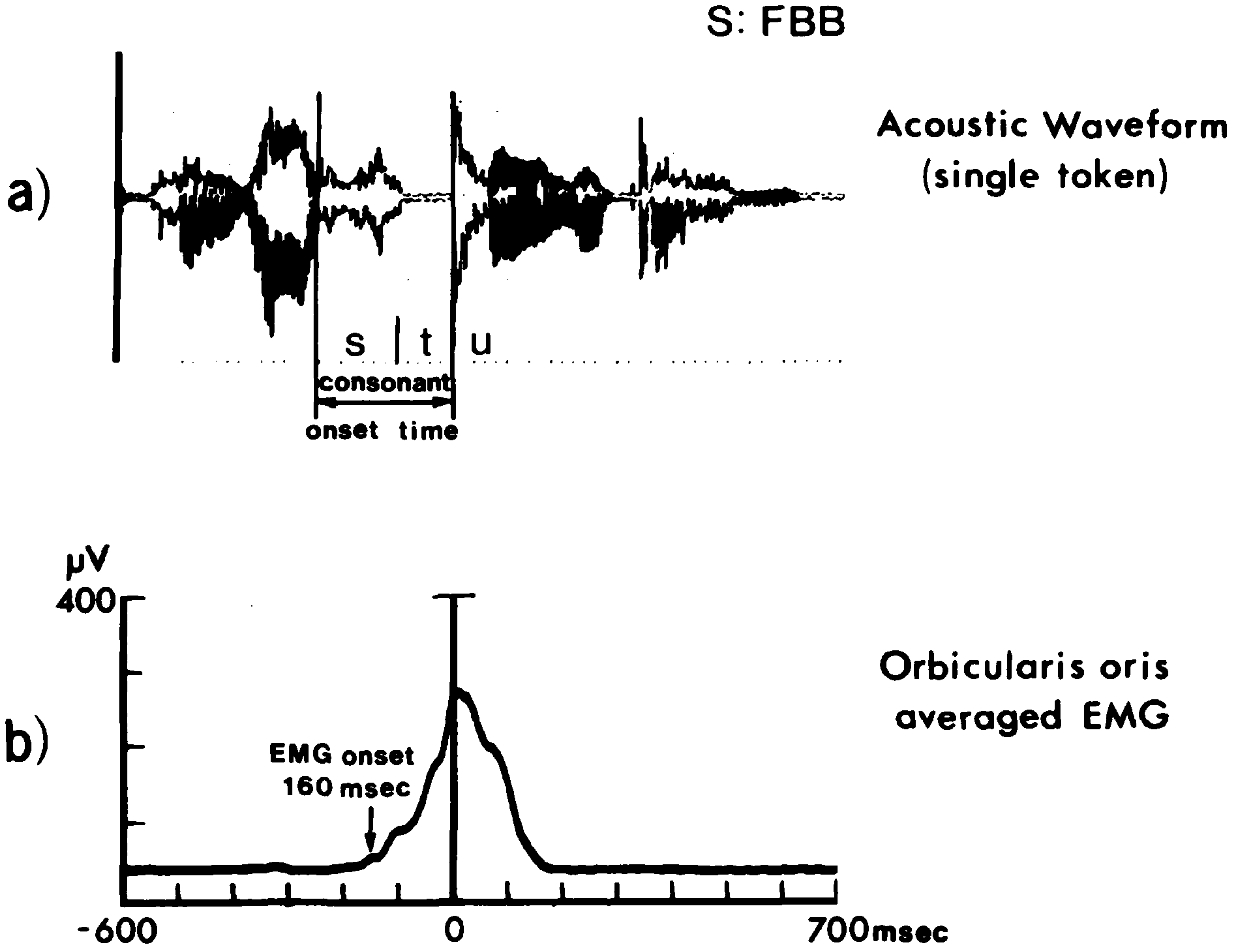


Figure 1. a) Sample waveform of one token of the utterance type "lease tool" for subject FBB, with consonant onset time (consonant string duration) indicated. b) Ensemble-average EMG activity from the orbicularis oris muscle, for all tokens of "lease tool," for subject FBB. EMG onset time, at 5% of baseline-to-peak amplitude, is indicated at 160 msec before /t/ release.

EMG and Audio Data Collection

EMG potentials were recorded from several placements on the superior and inferior orbicularis oris muscles for each subject, using surface electrodes similar to those described by Allen, Lubker, and Harrison (1972). The electrodes were applied to the vermillion border of the lips, and spaced about a half centimeter apart. The EMG signals were recorded simultaneously with the audio and clock signals on a multi-channel FM tape recorder. In later analyses, the channel yielding the EMG signal with the largest amplitude was chosen; in all cases, this was a superior lip placement. Signals from the lower lip placements did not appear to be qualitatively different, but had a lower signal-to-noise ratio.

Acoustic measurements. The acoustic recordings from each of the three subjects whose data were subjected to detailed analysis were digitized and analyzed using an oscillographic display of the digitized waveform. For each of the 18 two-word test utterances, the durations of the /lV/ and /Vl/ sequences were measured for each of the ten to eighteen repetitions, as were the durations of /s/ friction and /t/ closure and aspiration. Average durations of the /lV/, /Vl/, and consonant sequences were calculated from the individual token measurements.

Reference points were chosen for aligning tokens of each utterance type for later sampling and averaging of the EMG potentials. The point chosen for the 12 members of the anticipatory set was the release of the /t/ before /u/; for the carryover set, it was the moment of /t/ closure or the beginning of /s/ friction immediately after /u/ (Figure 1a).

EMG measurements. The EMG waveforms for each electrode position (channel) and utterance repetition were rectified, integrated (5 msec, hardware integration), and digitized. The signals were smoothed, using a 35-msec triangular window, and the ensemble average was calculated for each utterance and channel from the integrated EMG waveforms,[1] after aligning all tokens at the reference point in the acoustic waveform. These signal recording and processing techniques have been described in detail elsewhere (Kewley-Port, 1973).

Using the ensemble averages, we determined the beginning of orbicularis oris activity for the utterances in the anticipatory set, and the end of this activity for the utterances in the carryover set. For the anticipatory set utterances, the beginning of activity was defined as the time at which orbicularis oris EMG activity reached 5% of its maximum amplitude.[2] An example of an ensemble average of one utterance, from the data of subject FBB, is shown in Figure 1b, with this onset time indicated. For the carryover set, the end of activity was defined as the time at which orbicularis oris EMG activity fell to 5% of its maxi‑‑m amplitude.

RESULTS

Anticipatory Coarticulation

If the beginning of vowel rounding activity were linked to the beginning of the preceding consonant string, then, regardless of the number of conso-

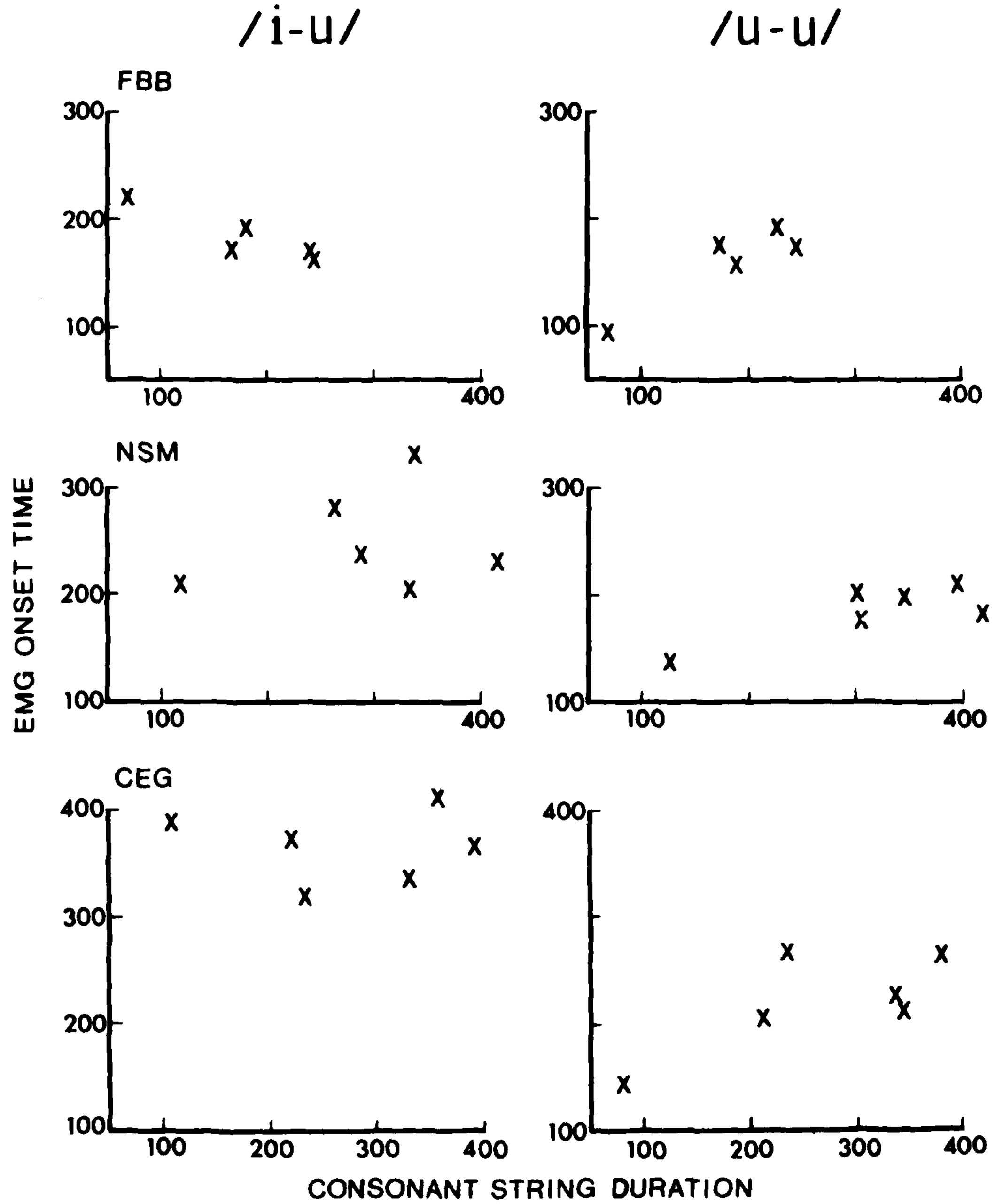


Figure 2. Scatter plots of consonant string duration vs. EMG onset time in msec for anticipatory set utterances, for all three subjects. /i-u/ utterance data are presented in the left-hand column; /u-u/ utterance data are presented in the right-hand column.

Table 1

A. Anticipatory Coarticulation: Slope of best-fit line for consonant string duration vs. EMG Onset Time for /i-u/ and /u-u/ utterances.

	FBB		NSM		CEG	
/i-u/	m=	-.3209	m=	.1049	m=	.0006
	*$F_{1,3}$=	19.49	$F_{1,4}$=	.20	$F_{1,4}$=	.000014
	m=	.4927	m=	.1953	m=	.2899
/u-u/	**$F_{1,3}$=	13.47	$F_{1,4}$=	7.21	$F_{1,4}$=	4.34

*$p<.05$, but slope is negative
**$p<.05$. If /utu/ case not included, m=.1544, $F_{1,2}$=.27 ($p>.05$).

B. Carryover Coarticulation: Slope of best-fit line for consonant string duration vs. EMG Offset Time for /u-i/ and /u-u/ utterances.

	FBB		NSM		CEG	
/u-i/	m=	.0161	m=	.0674	m=	.4295
	$F_{1,5}$=	.3249	$F_{1,4}$=	4.66	*$F_{1,4}$=	10.74
/u-u/	m=	-.0566	m=	.0162	m=	-.3843
	$F_{1,3}$=	.1089	$F_{1,4}$=	.2152	**$F_{1,4}$=	23.42

*$p<.05$
**$p<.01$, but slope is negative

nants in the string, this activity should begin earlier when the consonant string is of longer duration. If, on the other hand, the beginning of the orbicularis oris activity were linked to the presence of a rounded vowel, there should be no correlation between the timing of the beginning of EMG activity and the duration of friction and closure. Since there is a general tendency for these events to be of shorter duration in clusters, it is necessary to examine a number of different consonant sequences, of different lengths, in order to distinguish between the consonant-linked and vowel-linked onset hypotheses. In the present set, the acoustic durations of the medial sequences ranged from 70 msec to about 420 msec.

The "onset time" of orbicularis oris EMG activity relative to consonant-string duration is shown, for the utterances of the /i-u/ anticipatory set, in the left-hand column of Figure 2. Each panel shows the data for one of the three subjects; each point represents the average consonant-string duration and EMG onset time for about 14 tokens of each type for two subjects, and 10 tokens of each type for the third. If anticipatory coarticulation were systematically related to the onset of the consonant string, we would expect the points to be fitted by a line having a positive slope; instead, however, the points are fitted by a line whose slope is not significantly different from zero in two cases, and is significantly negative in the third (Table 1).

In the right-hand part of Figure 2, we have plotted the EMG onset time relative to consonant string duration for the /u-u/ utterances. The results fit the same general description as the /i-u/ case; that is, coarticulation began a constant interval before the onset of the second vowel, with a single exception for each of the three speakers--the case having the shortest consonant duration. A fairly straightforward explanation can be provided, if we assume that for this case the intervocalic interval may be shorter than the time necessary for muscle activity to fall to baseline for the first /u/ and rise for the second. This hypothesis is supported by the fact that, for all three subjects, the minimum, or baseline, activity for /t/ strings is higher than for any other (Table 2).

Table 2

Minimum EMG Amplitude (in Microvolts) During the Interval Between Vowels in /u-u/ Utterances

	FBB	NSM	CEG
t	68	79	114
#st	52	70	67
s#t	48	74	74
st#t	46	70	59
s#st	43	68	62
st#st		69	54

Another interesting result for the two vowel conditions is that there is a difference in the intercept of the best straight-line fit for /i-u/ and /u-u/ cases; that is, rounding for the second vowel begins earlier if the first vowel was /i/ than if it was /u/. Somewhat similar data are presented by Sussman and Westbury (1981), for /i-u/ sequences as contrasted with /a-u/ sequences. In their data, the difference in onset time is not significant for the /ikstu/ vs. /akstu/ comparison, although the difference in onset time is significant for the /iku/ vs. /aku/ comparison. If the differences in onset time are a consequence of the lip position for the first vowel, we might expect consistent amplitude differences for the second vowel, depending on the identity of the first. Such differences were reported by Sussman and Westbury for the /kst/ cases (see their Figure 3). They do not comment on the /k/ case, where one might expect larger effects. Peak EMG amplitudes for our own data are presented in Table 3, and, although there is some tendency for peak values for the second vowel to covary with the identity of the first, there is no absolutely consistent result.

The analysis presented in Figure 2 does not examine possible effects of the location of word boundaries. Indeed, in the classic experiment of Daniloff and Moll (1968), no effects of word boundaries were observed, although some similar experiments have claimed to show effects of some kinds of linguistic boundaries (e.g., McClean, 1973). Since there are complex but systematic effects of word boundaries on consonant duration (Lehiste, 1960), we re-examined the data for possible word-boundary effects, as shown in Figure 3. It was not possible to examine those utterances produced with a segment common to the end of the first word and the beginning of the second, because consonant duration could not be apportioned to one or another side of the word boundary. For example, as noted above, the sequence that was orthographically represented as "leased tool" was usually executed as [listul]; since /t/ was associated with both words, no separation could be made. For the subset of the utterances where an acoustic event could be associated with the word boundary, the results are as before--that is, there is no systematic relationship between onset of anticipatory coarticulation and word boundary (Figure 3). We would add that, for each utterance set for each subject, the range of EMG onset times for the orbicularis oris is considerably smaller than the range of consonant durations (Table 4, part A). If the onset of EMG activity were linked to the beginning of the measured durations, we would expect the ranges to be comparable.

Carryover Coarticulation

Examining the timing relationship between the end of orbicularis oris EMG activity and the duration of the consonant string following a rounded vowel, we found a pattern very much like that found for the anticipatory condition. Specifically, the "offset time" appears to be unaffected by the duration of the following consonant string (Figure 4). Rather, the slope of the line of best fit for each utterance set for each subject was not significantly different from zero (Table 1b). And, again as with the anticipatory coarticulation data, the range of EMG offset times is smaller than the range of consonant durations (Table 4, part B). In these data, however, lip position for the following vowel did not influence the timing of the end of the vowel gesture. That is, the following vowel is not anticipated in the timing of the end of the first vowel gesture.

Table 3

Peak EMG Amplitude (in Microvolts) for Vowels of Second Syllable of "Anticipatory" Set Utterances, with /u-u/ Utterance Peak Amplitude at the Left and /i-u/ Utterance Peak Amplitude at the Right

		peak amplitude	peak amplitude
		/u-u/	/i-u/
FBB			
	#t	236	362
	#st	314	301
	s#t	280	265
	st#t	239	269
	s#st	237	270
NSM			
	#t	518	439
	s#t	480	502
	st#t	475	444
	#st	506	507
	s#st	434	452
	st#st	421	430
CEG			
	t	272	222
	s#t	235	246
	#st	228	274
	s#st	207	200
	st#t	190	200
	st#st	240	244

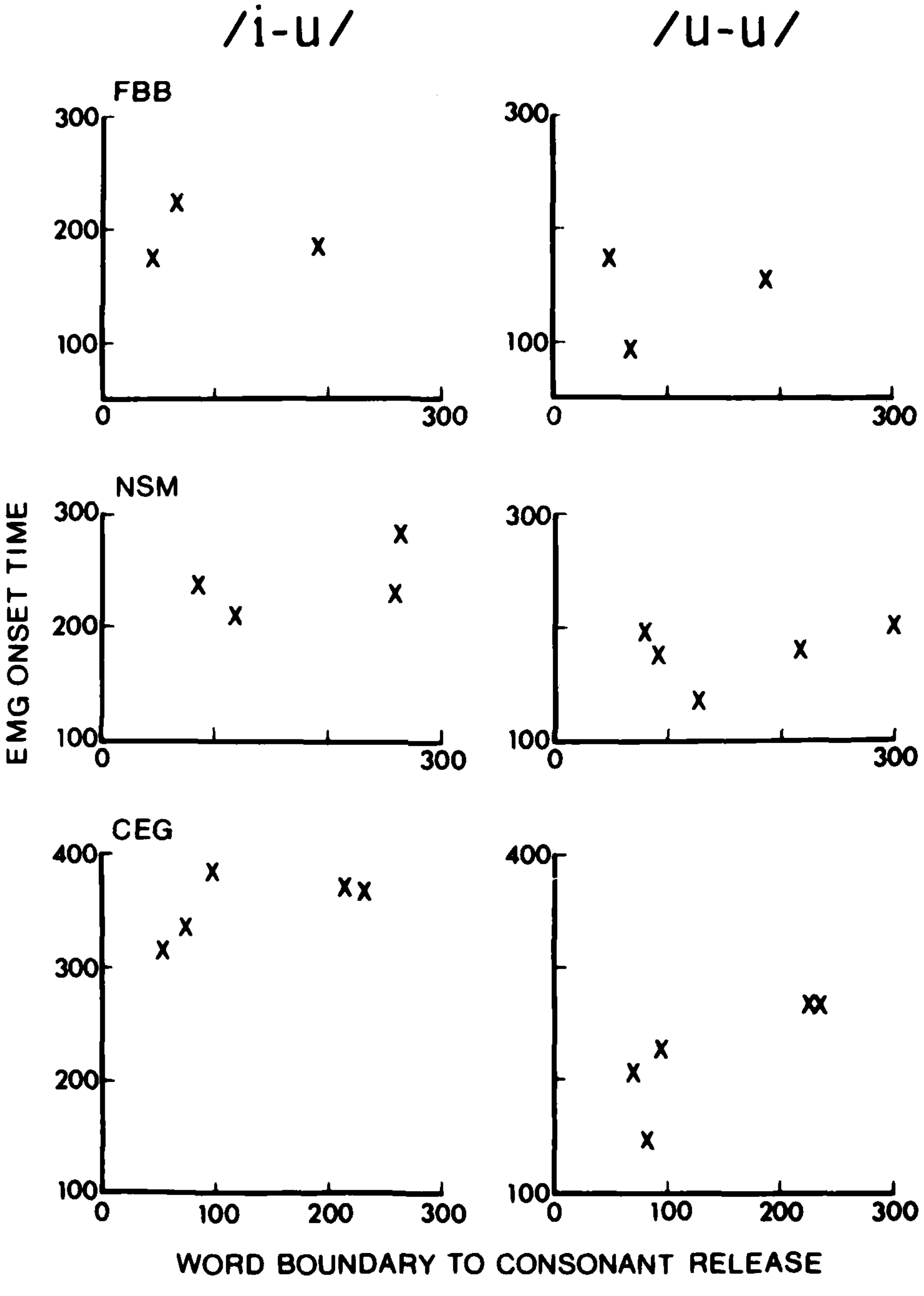


Figure 3. Scatter plots of the duration of word-initial consonant strings vs. EMG onset time in msec, for anticipatory set utterances, for all three subjects. /i-u/ utterance data are presented in the left-hand column; /u-u/ utterance data are presented in the right-hand column.

Table 4

Range, in Msec, of EMG Onset and Offset Times and Consonant String Durations

A. Anticipatory Coarticulation

		EMG Onset	Consonant Duration	Syllable Initial Consonant Duration
FBB	iC_nu	55	174	113
	uC_nu	95	172	119
NSM	iC_nu	125	299	176
	uC_nu	70	296	220
CEG	iC_nu	95	281	174
	uC_nu	120	298	166

B. Carryover Coarticulation

		EMG Offset	Consonant Duration	Syllable Final Consonant Duration
FBB	uC_ni	15	235	193
	uC_nu	50	172	123
NSM	uC_ni	25	293	211
	uC_nu	20	296	267
CEG	uC_ni	140	260	252
	uC_nu	110	298	244

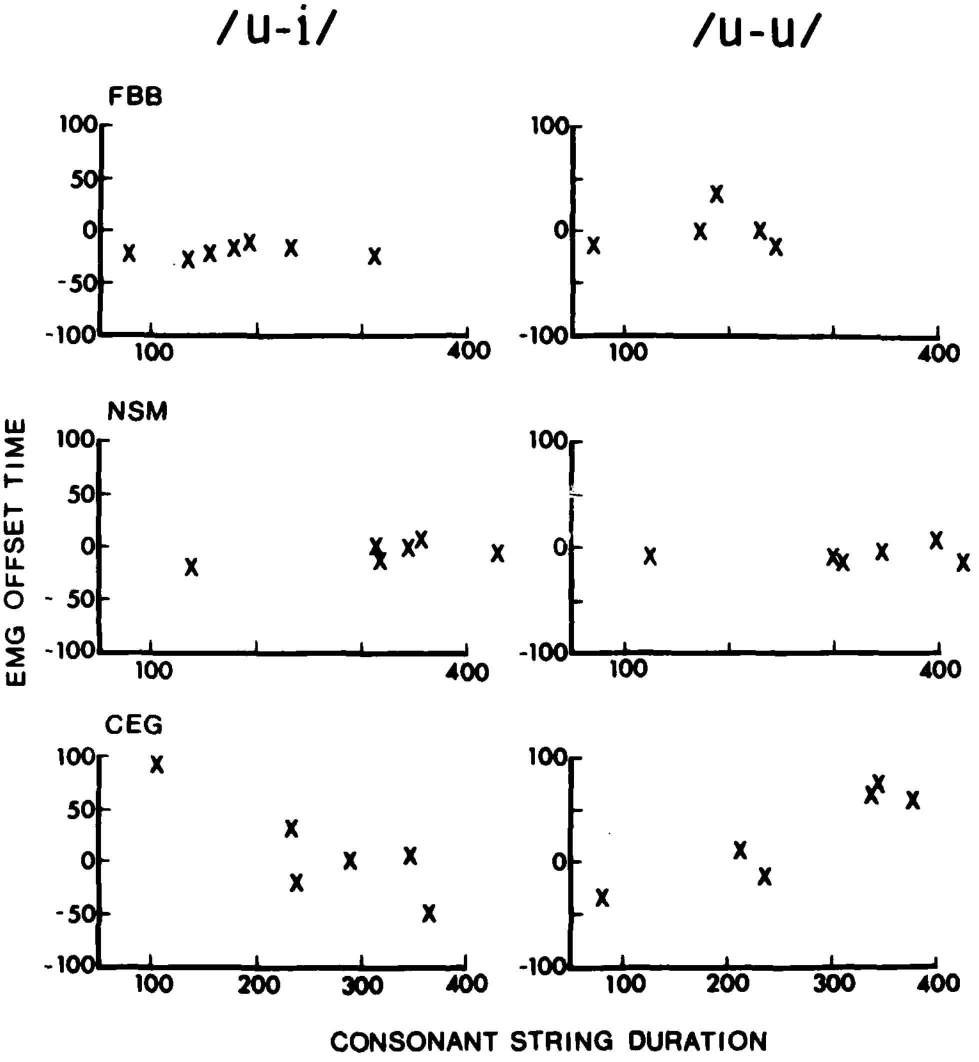


Figure 4. Scatter plots of consonant string duration vs. EMG offset time in msec, for carryover set utterances, for all three subjects. /u-i/ utterance data are presented in the left-hand column; /u-u/ utterance data are presented in the right-hand column.

DISCUSSION

The data suggest that the beginning of EMG activity associated with lip-rounding gestures for vowels is more obviously related to other components of the vowel articulation than to aspects of the consonant string length. Similarly, the end of EMG activity associated with lip-rounding gestures is most straightforwardly described with relation to the end of the vowel, and not with relation to the following consonant string.

Previously published reports, suggesting that lip-rounding gestures migrate ahead to the beginning of a preceding consonant string, may be accounted for by referring to the timing of orbicularis oris activity for the second vowel in /u-u/ utterances having short-duration consonant strings. In these cases, lip-rounding activity seems to begin later (i.e., closer to the second vowel) than it does in utterances having longer consonant sequences. If one examines only a few utterance types with one or two short and one long consonant sequence (cf. Sussman & Westbury, 1981), and if an earlier vowel gesture either inhibits or masks the beginning of the rounding gesture in the short-string utterances, it may appear as though lip-rounding onset follows the beginning of the preceding consonant string. However, we believe that our data cannot be accounted for in this way, nor can the movement study of Engstrand (1980), which give the same general picture.

This picture of coarticulation is quite different from the look-ahead scanner model, presented by Sussman and Westbury (1981). In their model, if a prior vowel is biomechanically antagonistic to rounding, "temporal and amplitude adjustments are incorporated into the anticipatory rounding gesture." Rounding begins, presumably, some time after the end of the antagonistic vowel, but this time is simply displaced, by some amount, from the beginning of the intervocalic string. Thus, there is always a carryover effect of the preceding vowel on the onset of rounding; but for all consonant strings longer than some value, the onset of rounding varies with string duration, presumably as a reflection of the number of elements in the string. In the model proposed here, a preceding vowel may have some antagonistic effect on the onset of rounding, and hence, rounding may appear closer to the second vowel in cases where the consonant string is short, or when the vowel changes. However, rounding onset time does not covary with the number of consonant string elements beyond that point. We assume that the reason Sussman and Westbury apparently observed a string-element effect is that they compared a one-consonant sequence with a three-consonant sequence.

There is still a good deal that remains unclear about both models and data. We agree that the onset of rounding is clearly influenced by peripheral biomechanical concerns; thus, in the Sussman and Westbury data, rounding for /u/ begins at a different time following /i/ and /a/, and, in our data, at a different time for /u/ following /u/ and /i/. However, by examining a set of utterances whose consonant durations for each subject were fairly well distributed through a wide range of durations, we believe we have shown the rounding gesture to be linked to the vowel articulation. That is, the specification of lip position for the consonants is not altered by a migrating vowel feature. Instead, and as we have also suggested elsewhere (Bell-Berti & Harris, 1981), we see the vowel-rounding gesture beginning at a relatively fixed time before the acoustic onset of the vowel and simply co-occurring with some portion of the preceding lingual consonant articulations.

REFERENCES

Allen, G. D., Lubker, J. F., & Harrison, E. New paint-on electrodes for surface electromyography. Journal of the Acoustical Society of America, 1972, 52, 124. (Abstract)

Baer, T., Bell-Berti, F., & Tuller, B. On determining EMG onset time. In J. J. Wolf & D. H. Klatt (Eds.), Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America. New York: Acoustical Society of America, 1979.

Bell-Berti, F. Velopharyngeal function: A spatial-temporal model. In N. J. Lass (Ed.), Speech and language: Advances in basic research and practice (Vol. 4). New York: Academic Press, 1980.

Bell-Berti, F., & Harris, K. S. More on the motor organization of speech gestures. Haskins Laboratories Status Report on Speech Research, 1974, SR-37/38, 73-77.

Bell-Berti, F., & Harris, K. S. Anticipatory coarticulation: Some implications from a study of lip rounding. Journal of the Acoustical Society of America, 1979, 65, 1268-1270.

Bell-Berti, F., & Harris, K. S. A temporal model of speech production. Phonetica, 1981, 38, 9-20.

Benguerel, A.-P., & Cowan, H. A. Coarticulation of upper lip protrusion in French. Phonetica, 1974, 30, 41-55.

Bronstein, A. J. Pronunciation of American English. New York: Appleton-Century-Crofts, 1960.

Daniloff, R. G., & Moll, K. L. Coarticulation of lip rounding. Journal of Speech and Hearing Research, 1968, 11, 707-721.

Engstrand, O. Acoustic constraints or invariant input representation? An experimental study of selected articulatory movements and targets. Reports from Uppsala University Department of Linguistics, 1980, 7, 67-95.

Fowler, C. A. Coarticulation and theories of extrinsic timing control. Journal of Phonetics, 1980, 8, 113-133.

Fromkin, V. A. Neuromuscular specification of linguistic units. Language and Speech, 1966, 9, 170-199.

Gay, T. J. Articulatory units: Segments or syllables?. In A. Bell & J. B. Hooper (Eds.), Syllables and segments. Amsterdam: North Holland Publishing Company, 1978.

Harris, K. S., Lysaught, G. F., & Schvey, M. M. Some aspects of the production of oral and nasal labial stops. Language and Speech, 1965, 8, 135-147.

Henke, W. L. Dynamic articulatory model of speech production using computer simulation. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1966.

Kewley-Port, D. Computer processing of EMG signals at Haskins Laboratories. Haskins Laboratories Status Report on Speech Research, 1973, SR-33, 173-183.

Lehiste, I. An acoustic-phonetic study of internal open juncture. Phonetica, 1960, 5(Supplement), 1-54.

MacNeilage, P. F. Motor control of the serial ordering of speech. Psychological Review, 1970, 77, 182-196.

McClean, M. Forward coarticulation of velar movement at marked junctural boundaries. Journal of Speech and Hearing Research, 1973, 16, 286-296.

Sussman, H. M, & Westbury, J. R. The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. Journal of Speech and Hearing Research, 1981, 22, 16-24.

Tatham, M. A. A., & Morton, K. Some electromyography data towards a model of speech production. University of Essex Language Centre Occasional Papers, 1968, 1, 1-59.

FOOTNOTES

[1] Optimum choice of timing measures from EMG signals depends on several considerations, including both the nature of the EMG data themselves and the use for which the measurements are intended. There are three sources of token-to-token variability in EMG signals whose relative magnitudes bear on the choice: uncorrelated electrical noise, the statistical nature of motor-unit excitation, and articulatory timing variation. Effects of this third source are minimized by control of speaking rate and by judicious choice (and careful measurement) of the acoustic reference point. When the first two sources of variability are large--and especially when the EMG onsets are gradual--measurement from the average signal is preferred. Since we frequently encounter both gradual onsets and relatively noisy signals, use of the ensemble average in determining EMG onset time is generally the method of choice (Baer, Bell-Berti, & Tuller, 1979).

[2] This value was chosen because it assured that we were not identifying random background noise as the beginning of activity. This 5% point was exceeded for each speaker for the utterance "loo tool," which had a relatively short "consonant string" and, consequently, the minimum level of EMG activity between the two rounded vowels did not fall to 5% of the peak activity. For these cases, we chose the times at which minimum activity occurred.

TEMPORAL CONSTRAINTS ON ANTICIPATORY COARTICULATION*

Carole E. Gelfer,+ Katherine S. Harris,+ and Gary Hilt++

Abstract. Two accounts of coarticulation, 1) that the anticipation of segmental gestures, and thus the extent of their influence, is determined primarily according to the compatibility of the feature specifications for preceding and anticipated phones, and 2) that the extent of anticipatory gestures is delimited according to temporal specifications intrinsic to the motor program, yield very different predictions regarding articulatory organization. These predictions were tested by varying the number of intervocalic consonants in a V1CnV2, where V2 was either /i/ or /u/ and Cn was /s/, /st/, or /st#st/. We were thus able to determine the extent of spectral changes within the consonant string as a function of the upcoming vowel. Our results lend support to the second account and suggest that the onset of a phone's influence on preceding segments is temporally constrained, presumably because anticipatory gestures are time-locked to the segments they characterize and are not freely-migrating features.

A significant issue in speech production theory is the extent to which articulatory gestures for speech segments are anticipated. From the long-ago realization that phones were, at least spectrographically, nondiscrete, theories of feature spreading were born in attempts to reconcile a continuous output with a presumed noncontinuous input (e.g., Daniloff & Hammarberg, 1973; Henke, 1967).

Numerous models of coarticulation have incorporated the notion that the anticipation of articulatory gestures occurs primarily according to the compatibility of the feature specifications for preceding and anticipated phones (Benguerel & Cowan, 1974; Daniloff & Moll, 1968; Henke, 1967; McClean, 1973; Sussman & Westbury, 1981). Coarticulation, according to this view, is therefore limitless with regard to time and spreads over entire phonological units until it is blocked by incompatible gestures. In anticipation of a rounded vowel, for example, lip rounding is said to occur over as many previous segments as are unspecified for lip configuration, with the extent of the anticipatory gesture varying directly with the onset of the prevocalic string (see, for example, Benguerel & Cowan, 1974; Daniloff & Moll, 1968).

*A version of this paper was presented at the 101st Meeting of the Acoustical Society of America, May 18-21, 1981, Ottowa, Canada.

+Also The Graduate School, City University of New York.

++Also The Graduate School, City University of New York and The Veterans Administration Medical Center, Bronx, New York.

Acknowledgment: This work was supported by NINCDS Grants NS-13617 and NS-13870 and BRS Grant RR-05596 to Haskins Laboratories.

An alternative to segment-based models of coarticulation is one that posits that anticipatory gestures are time-locked to the segments they characterize (Bell-Berti, 1980; Bell-Berti & Harris, 1979, 1981; Fowler, 1980). Such a model would also predict the coproduction of a larger number of segments as a consonant string increases, not because an anticipatory gesture attaches itself to all preceding phones, but as a result of segmental shortening. Thus, an increasing number of segments fall within the relatively fixed time course of the articulatory gesture. The extent of anticipation is therefore temporally delimited, without regard to the absolute number of preceding segments.

A temporal model further predicts that the magnitude of an up-coming phones's influence should vary as a function of temporal proximity to that phone. Therefore, the longer the preceding string, the less likely it is to show coarticulatory effects at its onset, while a shorter string should show effects over proportionately more of its length. However, at the same point in time relative to the acoustic onset of the up-coming phone, the degree of the upcoming vowel's influence should be similar for both long and short strings.

Alternatively, if, as segment-based models posit, the onset of coarticulation occurs simultaneously with the onset of a preceding string, we would expect anticipatory effects to be determined by a segment's proximity to the onset of the string and to bear little or no relation to its temporal distance from the second vowel.

Procedure

The predictions of either model should be verifiable by varying the number of items in intervocalic consonant strings and noting the pattern of second formant frequency changes as a function of distance from the second vowel. We therefore constructed real-word utterances with VCV's embedded, where C was /s/, /st/, or /st#st/ and V_1 and V_2 were either /i/ and /u/ alternately or both /i/ or /u/. In all there were twelve utterance types or six minimal pairs, differing only in the identity of V_2. Two male speakers of metropolitan New York area dialect read 20 repetitions of each utterance, embedded in the carrier phrase "Not ________ today" (e.g., "Not lease ease today"). Productions were recorded on magnetic tape, input to a Honeywell DDP-224 interactive computer at Haskins Laboratories, and digitized. Token waveforms were aligned to the onset of the second vowel and spectrum-analyzed. Line-up points were confirmed spectrographically by noting a sudden increase in F_1 amplitude at identical spectral sections.

Because we were interested in making frequency measurements within the consonant string, and therefore within friction, individual spectra for each utterance were computer-averaged in order to increase the reliability of our measurements. We were thus able to observe and measure low frequency resonances within the friction, and we will refer to these as second formants because they appear to be continuous with the second formants of the flanking vowels. Soli (1981) and Yeni-Komshian & Soli (1979) have also consistently noted such low frequency resonances within friction for single instances of fricatives produced both in isolation and in vocalic environments.[1]

Figure 1 shows two averaged spectra with their respective peaks displayed above for the minimal pair "lease ease" and "lease ooze" for one subject. Note the low frequency resonance through the intervocalic portion of these utterances.

Figure 2 shows the averaged waveforms for the utterances "lease ease," "beast ease" and "least steel" for a second subject. These are 200 msec samples that include the first 50 msec of the second vowel and the 150 msec preceding it. Thus, at every temporal point relative to the onset of V_2, we are sampling a different portion of the acoustic signal for each utterance type.

It should also be noted that, for the /st#st/ utterance, despite the orthography, there is evidence of only one friction portion, one closure period and one release. Thus, this utterance appears to have been produced "naturally," that is, as [st:], differing from the /st/ utterance only in the duration of the closure.

While spectral averaging solves some problems, however, it also presents others. Thus, because individual tokens of a given utterance type are produced with variable durations, it is likely that the friction and vocalic portions will be averaged together as the distance from the second vowel increases. In order to minimize the possibility of confounding the data in this way, we took the range for all tokens of each consonant string type, determined the midpoint, and sorted tokens into long and short bins on this basis.

F_2 measurements were made from spectral sections at 12.5 msec intervals and collapsed over 25 msec intervals for the 150 msec preceding the acoustic onset of the second vowel. For each minimal pair, F_2 values for the utterances with final /u/ were subtracted from those of utterances with final /i/. Since initial vowels were always identical, positive values are therefore indicative of the final vowel's influence, with larger differences reflecting greater anticipatory effects.

Results

Figure 3 shows the difference in Hz along the y-axis for F_2 for all long and short minimal pairs where V_1 is /i/ after sorting. It should be noted that, when tokens are sorted in this way, there is temporal overlap between utterance types. For example, the longest singleton string is longer than the shortest /st/ string, while the longest /st/ strings are comparable in duration to the shortest /st#st/ strings, which, it should be recalled, were pronounced [st:]. Thus, these figures actually depict two--and sometimes three--comparisons: one for consonant strings of different phonetic structure and duration, one for consonant strings of identical phonetic structure but different durations, and, in some cases, one where phonetic structure differs but durations are comparable.

What the data show is that, despite temporal and phonetic differences or similarities, the critical variable appears to be time from the onset of the second vowel, such that there is a similar decrease in the F_2 difference for each pair as their distance from V_2 increases. In other words, it appears that, for utterances of this type, the influence of the second vowel is

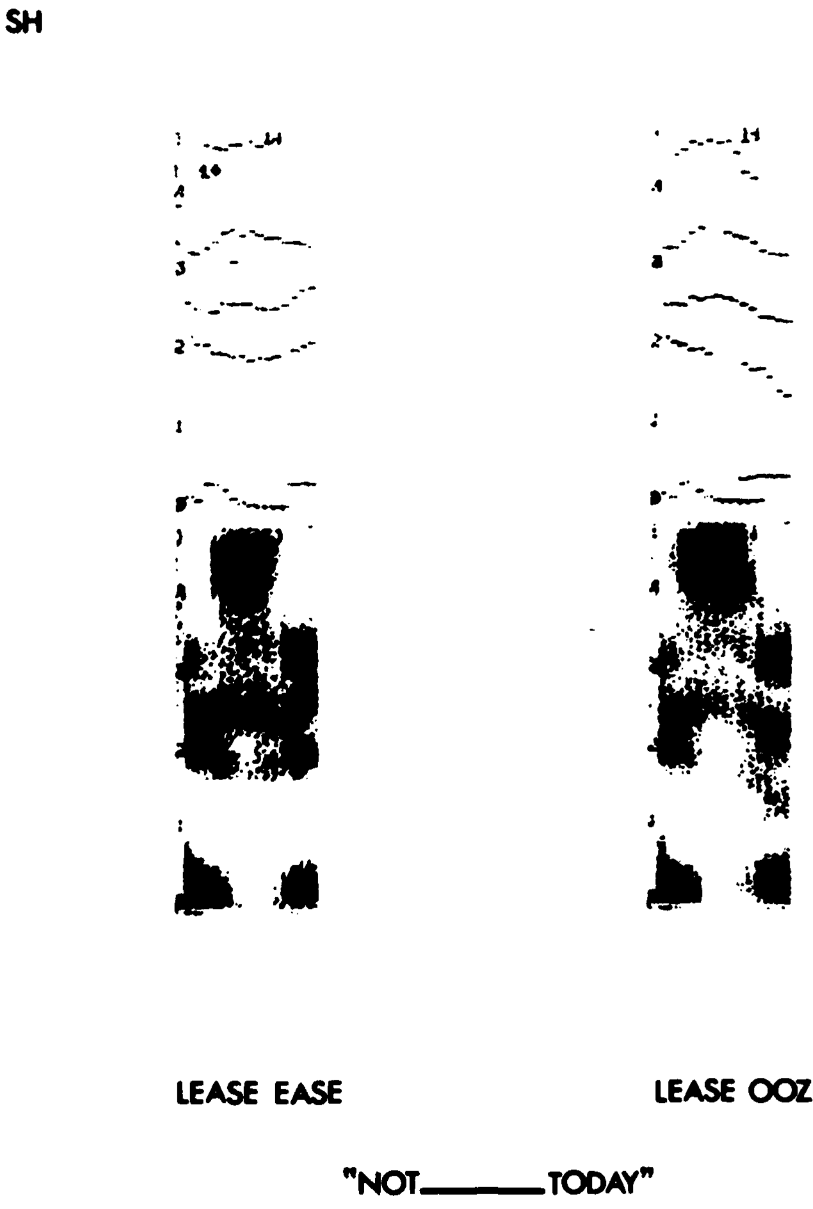


Figure 1. Two averaged spectra with their respective peaks displayed above for the minimal pair "lease ease" and "lease ooze" for one subject.

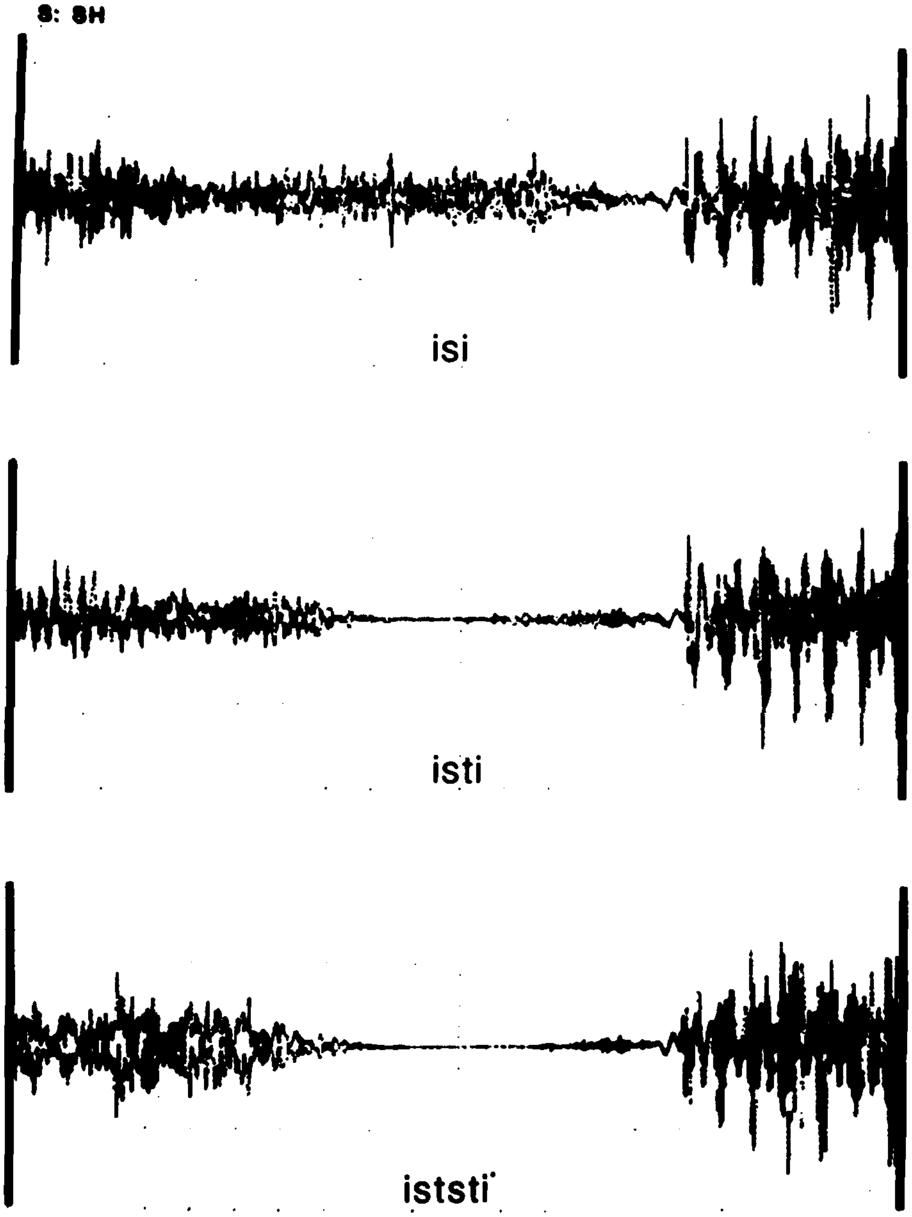


Figure 2. Averaged waveforms for the utterances "lease ease," "beast ease" and "least steel" for one subject. Accompanying labels depict only the intended expressions and are not transcriptions of the subjects' actual productions.

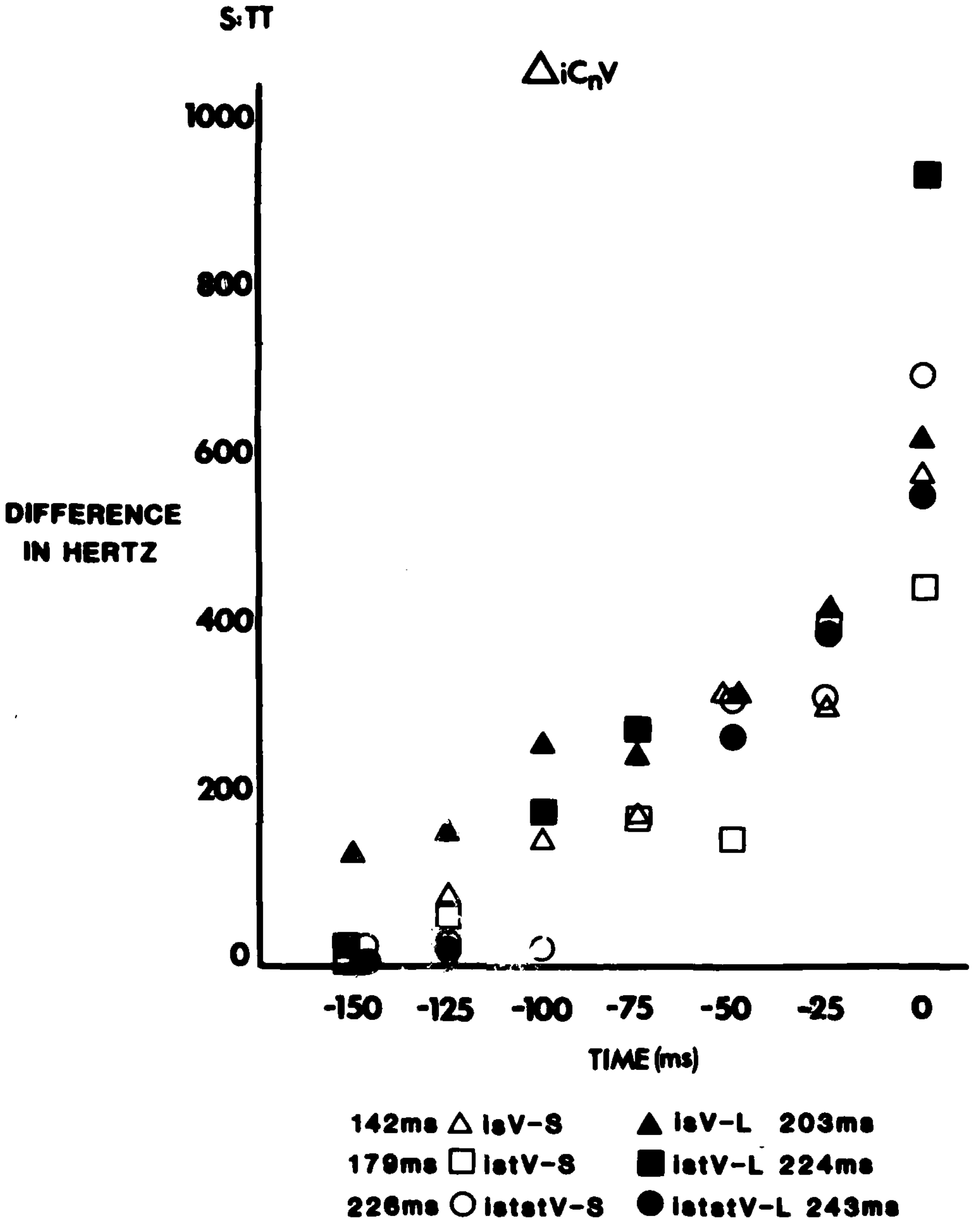


Figure 3. F2 difference in Hz for sorted tokens of minimal pairs where V1 is /i/. Long and short tokens are indicated by closed and open symbols, respectively. The different symbols denote consonant string type, with triangles for singleton intervocalic strings, squares for /st/ strings, and circles for /st#st/ strings. Values on the x-axis indicate time before the onset of V2, which is indicated by 0. Temporal points where symbols are absent correspond to the closure period of the stop consonant. The msec values next to the symbols in the legend indicate average consonant string durations for each minimal pair.

temporally delimited irrespective of the segmental composition of the preceding string.

Figure 4 depicts the same F_2 difference as a function of time from the onset of the second vowel, but for pairs where V_1 is /u/. Again, the long and short tokens of mimimal pairs are plotted, and there is the same temporal overlap for tokens of different phonetic structure. Perhaps even more than the first figure, the data illustrate the tendency for all utterance types to show similar anticipatory effects at almost all sampled intervals.

Note, too, that at -150 msec we are sampling the F_2 difference at the end of the first vowel for the shortest /s/ tokens. It is interesting that the magnitude of this difference is almost identical with that of the friction portion of the other pairs. This finding might be explained, not by anticipatory lip configurations as far back as the first vowel, which for /i/ and /u/ are incompatible, but by tongue configurations that are capable of anticipating up-coming phones without preventing the successful production of current ones. Thus, the job of coproduction may be divided between primary articulators.

Figure 5 shows the data for our second subject's minimal pairs where V_1 is /i/. While the trend is similar in the sense that anticipatory effects are similar in magnitude at most intervals, the effects diminish more abruptly over time and at intervals closer to V_2.

A possible explanation is the fact that, with only the exception of the /st#st/ pairs, all V_1 offsets occur within this 150 msec window. This is unlike our first subject, whose consonant strings were of longer durations and, with one exception, fell outside this time frame. Thus, while it may be possible for these vowels to coarticulate, and therefore show anticipatory effects, there may be limits to these effects for vowels as opposed to friction, thus possibly accounting for the rapid fall-off in F_2 differences.

It is interesting, too, that there are some negative values, indicating a higher F_2 when /u/ rather than /i/ is the second vowel. However, almost all of these occur at 150 msec prior to the acoustic onset of V_2, the most remote portion of our sample. And, while we have not tested these differences statistically, we would speculate that most of these values do not deviate significantly from zero. The value for the long /st#st/ pair, however, is at approximately minus 100 Hz, which is substantial, if not significant. And, since there is no other instance of such a negative value, it is possible that this reflects carry-over effects.

Figure 6 shows the data for the second subject's pairs where V_1 is /u/, and it is similar to his other utterances in that there is an abrupt fall-off in magnitude of the F_2 difference at -75 msec. The general trend is, however, similar, although there is more scatter at the intervals farthest from V_2, which we cannot explain. This differs not only from our other speaker, but also from this speaker's other utterances.

Discussion

The data for both subjects show the tendency for coarticulatory effects to be maximal at points in time closest to the acoustic onset of the second

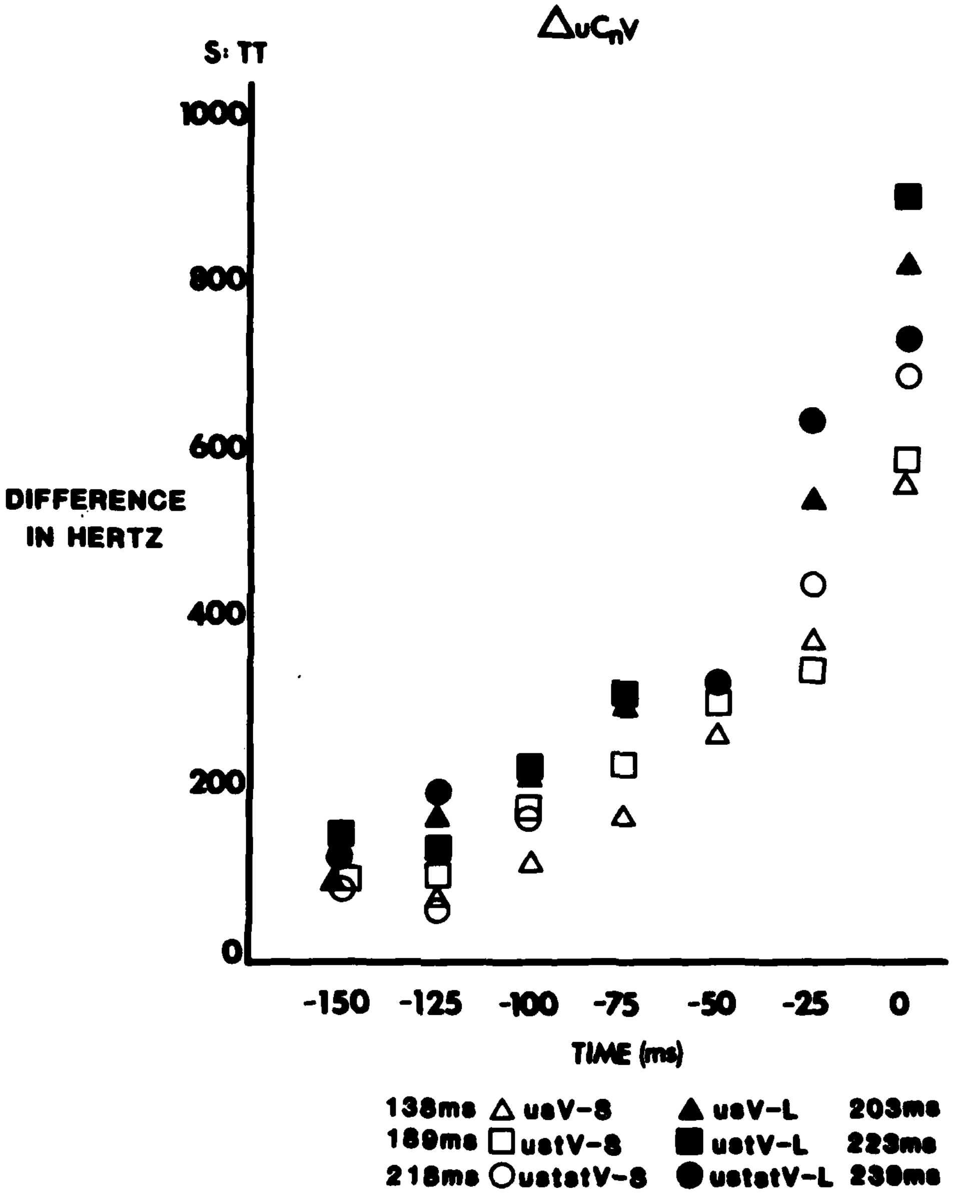


Figure 4. F2 differences in Hz for sorted tokens of minimal pairs where V1 is /u/.

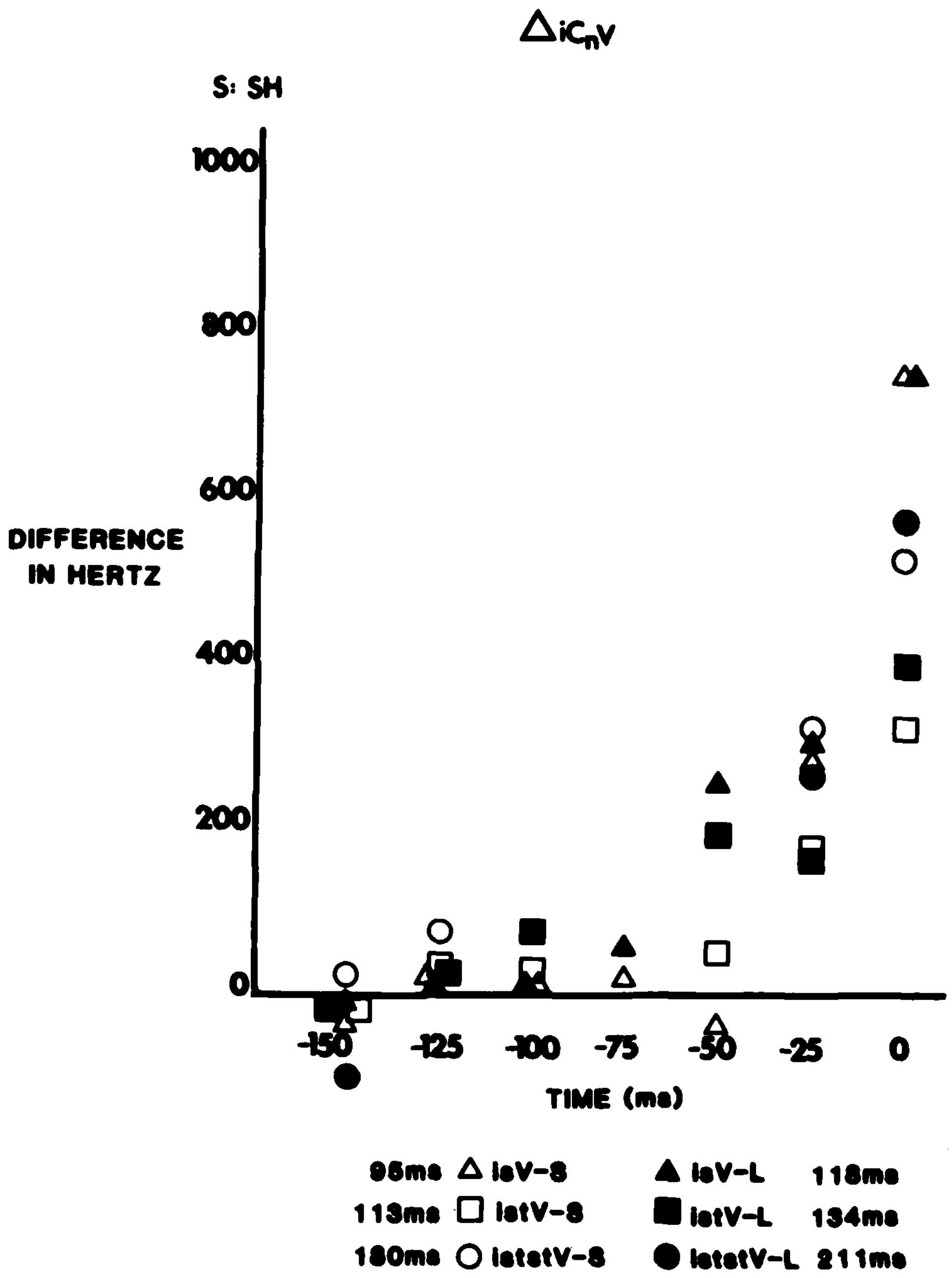


Figure 5. F2 differences in Hz for sorted tokens of minimal pairs where V1 is /i/ for second subject.

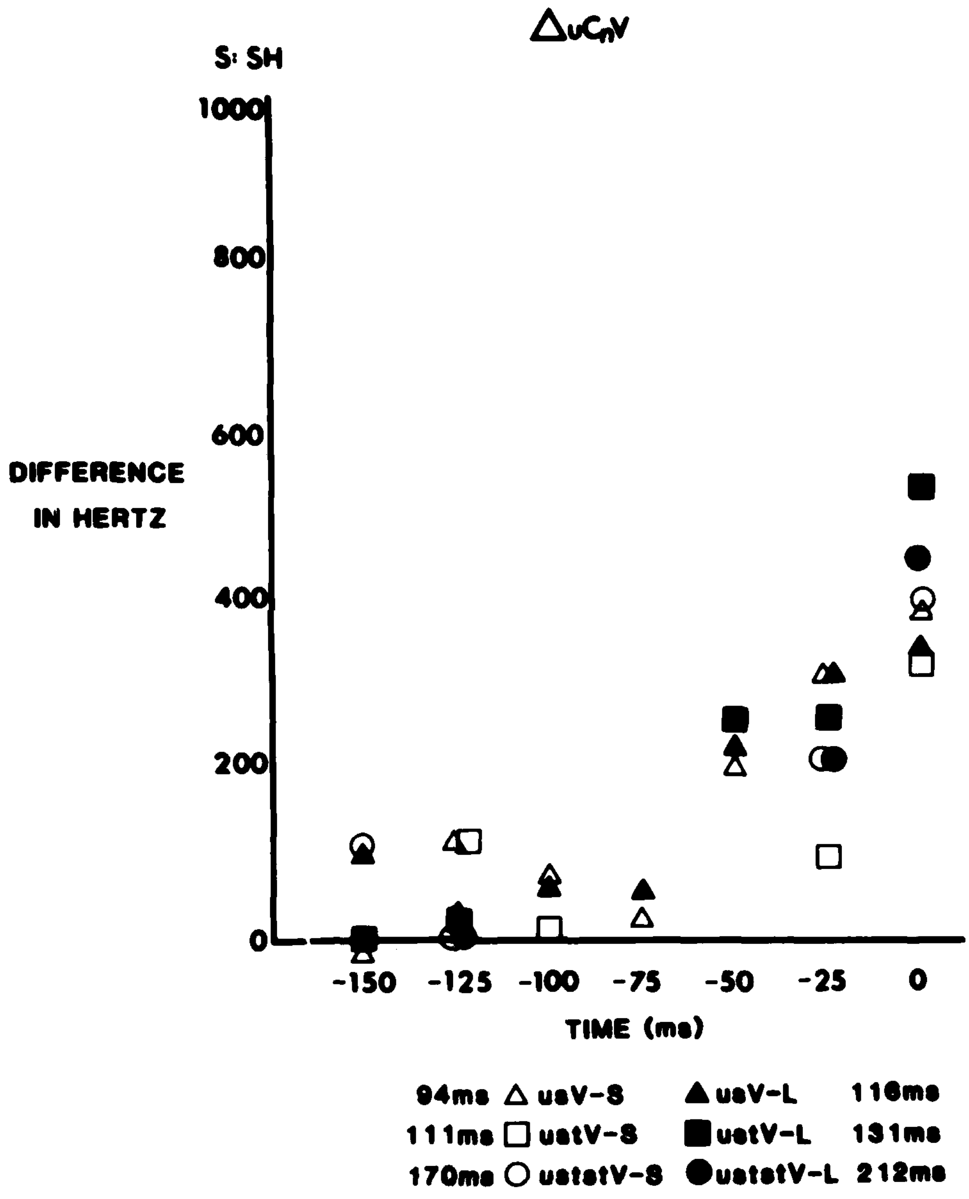


Figure 6. F2 differences in Hz for sorted tokens of minimal pairs where V1 is /u/ for second subject.

vowel, independent of absolute duration and segmental composition of the preceding consonant string. And, while we do not observe the influence of V_2 to be identical in magnitude at all points in time, the effects are systematic enough to support the notion that coarticulation is temporally constrained.

The data thus speak against the notion that anticipatory gestures automatically extend back to the onset of a preceding string. It was observed that the early portions of the longer strings failed to show substantial effects of the second vowel even though they were allegedly free to do so in the sense that anticipation of V_2 was in no way incompatible with their successful production. Furthermore, some of these F_2 differences were actually reversed, indicating, perhaps, that carry-over effects were still operative during the early portion of these strings. In addition, coarticulatory effects for the shortest consonant strings were sometimes observable during the latter portion of the first vowel. Thus, we see both the absence of coarticulatory effects in places where segment-based models predict their occurrence, as well as the presence of effects where these models, by virtue of the hypothesized mechanisms, predict their absence.

Our acoustic data are consistent with those of Soli (1981), who found the frequency of F2 within friction to be lower in anticipation of /u/ vs. /i/. However, he attributes this difference, not to lip rounding, but to different place of the primary constriction in anticipation of back vs. front vowels. His argument appears to derive primarily from data showing F2 frequencies to be similar preceding /a/ and /u/, where both are back vowels but only one is rounded. According to Soli, the effect of rounding, then, is to alter the fricative's overall spectral shape above 3 kHz. He maintains further that "while anticipatory vowel coarticulation appears to be limited to the final portion of the fricative," anticipatory lip rounding may occur throughout the fricative (p. 21).

While we consider Soli's general hypothesis regarding the acoustic effects of anticipatory tongue configurations to be a very tenable one, we would reject the notion that the general time course of anticipatory gestures differs significantly for different articulators. In other words, the fact that the lips are free to round during the course of a fricative preceding /u/ does not mean that they do so. This was demonstrated electromyographically by Bell-Berti and Harris (1979, in press) and cineradiographically by Engstrand (1981), whose data show lip rounding to occur at a fixed time before the acoustic onset of a rounded vowel and to be unaffected by the number of preceding consonant segments, the production of which in no way precluded lip rounding. In addition, Bell-Berti and Harris (in press) demonstrated that certain speakers round for /s/ in totally unrounded environments (e.g., /isi/). Thus, one would naturally expect the electromyographic and acoustic records to differ depending on whether rounding is or is not an inherent feature of a speaker's fricative production.

The main point here is that while it may be that lip rounding and place of constriction exert different spectral influences, it is intuitively unreasonable as well as empirically unfounded to suppose that the general organization of anticipatory gestures should be articulator-specific.

The results of the present study suggest that the onset of a vowel's influence on preceding segments is temporally constrained, presumably because anticipatory gestures are time-locked to the segments they characterize as opposed to being freely-migrating features. Further interpretation of the data, however, is limited by the fact that only the acoustic waveform was analyzed. We are currently planning studies with simultaneous EMG recordings from orbicularis oris and pertinent intrinsic and extrinsic tongue musculature in order to determine whether we can account for our acoustic data and Soli's on the basis of tongue and/or lip configurations. In addition, using subjects who produce /s/ with and without rounded lips in nonrounded envirnonments should provide an interesting comparison.

REFERENCES

Bell-Berti, F. Velopharyngeal function: A spatial-temporal model. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 4). New York: Academic Press, 1980, 291-316.

Bell-Berti, F., & Harris, K. S. Anticipatory coarticulation: Some implications from a study of lip-rounding. *Journal of the Acoustical Society of America*, 1979, *65*, 1268-1270.

Bell-Berti, F., & Harris, K. S. A temporal model of speech production. *Phonetica*, 1981, *38*, 9-20.

Bell-Berti, F., & Harris, K. S. Temporal patterns of coarticulation: Lip Rounding. *Journal of the Acoustical Society of America*, in press.

Benguerel, A-P., & Cowan, H. A. Coarticulation of upper lip protrusion in French. *Phonetica*, 1974, *30*, 41-45.

Daniloff, R. G., & Hammarberg, R. E. On defining coarticulation. *Journal of Phonetics*, 1973, *1*, 239-248.

Daniloff, R. G., & Moll, K. L. Coarticulation of lip-rounding. *Journal of Speech and Hearing Research*, 1968, *11*, 707-721.

Engstrand, O. Acoustic constraints or invariant input representation? An experimental study of selected articulatory movements and targets. *Ruul 7* (Reports from Uppsala University, Department of Linguistics), 1981, *nr 7*, 67-95.

Fowler, C. A. Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, 1980, *8*, 113-133.

Heinz, J. M., & Stevens, K. N. On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, 1961, *33*, 589-596.

Henke, W. Preliminaries to speech synthesis based on an articulatory model. *Conferences Preprints: 1967*. Conference on Speech Communication and Processing (Air Force Cambridge Research Laboratories, Bedford, Massachusetts), 1967, 170-177.

McClean, M. Forward coarticulation of velar movement at marked junctural boundaries. *Journal of Speech and Hearing Research*, 1973, *16*, 286-296.

Soli, S. Second formants in fricatives. *Journal of the Acoustical Society of America*, 1981, *69*, S5. (Abstract)

Sussman, H. M., & Westbury, J. R. The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory labial coarticulation. *Journal of Speech and Hearing Research*, 1981, *24*, 16-24.

Yeni-Komshian, G., & Soli, S. D. Extraction of vowel information from fricative spectra. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers presented at the 97th Meeting of the Acoustical Society of America*. New York: Acoustical Society of America, 1979, 37-40.

FOOTNOTES

[1] It should be noted that while we and others (Yeni-Komshian & Soli, 1979; Soli, 1981) consistently note low frequency resonances within friction, previous accounts of the acoustic theory of fricative production (e.g., Heinz & Stevens, 1961) all but dismiss the presence of low frequency resonances, due either to the decoupling of the front and back cavities or to the cancellation of back cavity resonances by the presence of zeroes.

IS A STOP CONSONANT RELEASED WHEN FOLLOWED BY ANOTHER STOP CONSONANT?*

Janette B. Henderson+ and Bruno H. Repp

Abstract. Many phonetics textbooks state that, in sequences of two stop consonants in English, the first stop is commonly unreleased. For nonhomorganic stop consonant sequences, this statement may be taken to imply that the (necessary) articulatory release of the first stop has no observable acoustic consequences. To examine this claim, we recorded sentences, produced by several native speakers of American English at a conversational rate, containing word-internal sequences of two nonhomorganic stops, either across a syllable boundary (e.g., cactus, pigpen), or in word-final position (e.g., act, sobbed). Oscillograms of the critical words revealed that release bursts of the first stop occurred in the majority of tokens, except in those where the second stop was bilabial. The bursts were acoustically rather weak and difficult to detect by ear, which may account for their having been neglected in the literature. Instead of a simple "released"-"unreleased" distinction, we propose a five-way classification that makes use of articulatory, acoustic, perceptual, and contrastive phonetic criteria.

INTRODUCTION

In English, sequences of two nonhomorganic stop consonants are not uncommon. They occur across word boundaries (e.g., big dog, great game), across syllable boundaries within words (e.g., cactus, pigpen), and in word-final position (e.g, act, sobbed). Textbooks of English phonetics generally point out that the first stop in such sequences is commonly unreleased or unexploded. Some authors (e.g., Ladefoged, 1975, pp. 45, 49; MacKay, 1978, p. 166) say no more than that, while others (e.g., Abercrombie, 1967, p. 146; Catford, 1977, p. 222; Jones, 1956, p. 155; Kenyon, 1951, p. 47) are more explicit about the articulatory and acoustic events involved.

Without further qualification, the statement that the first stop in a two-stop sequence is unreleased may be misleading. If "release" is correctly interpreted as a strictly articulatory term, referring to the breaking of contact between two articulators that results in the release of overpressure built up behind the occlusion, the statement obviously cannot be true if the two stops have different places of articulation. In sequences of two

*To appear in Phonetica. This is a revised and expanded version of a paper presented at the 101st Meeting of the Acoustical Society of America in Ottawa, Ontario, 18-22 May 1981. The research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to Haskins Laboratories. We thank Arthur Abramson, Leigh Lisker, and Ignatius Mattingly for helpful comments on an earlier draft.

+Also Department of Linguistics, University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-67/68 (1981)]

nonhomorganic stops, the closure of the first stop must be released before that of the second stop; otherwise, the second stop would be produced with an incorrect or dual place of articulation. Therefore, it appears that phoneticians have used the terms "release" and (perhaps less ambiguously) "explosion" to refer not to the articulatory release but to its acoustic consequences--the portion of the speech signal that, for reasons of terminological consistency (cf. Repp, 1981), we prefer to call the "release burst." If so, then a strict interpretation of the term "unreleased" would imply that, even in sequences of two nonhomorganic stop consonants, no release burst of the first stop is found in the acoustic record.

We were surprised, therefore, when earlier measurements of VCCV nonsense utterances (where CC was a two-stop sequence) produced by three speakers revealed that the large majority of the tokens contained clearly identifiable release bursts of the first stop (Repp, 1980). In a more recent study using similar utterances produced by two speakers (Repp, in press), all tokens (with a single exception) contained such bursts; moreover, the bursts were shown to have perceptual significance. Were earlier authors wrong, or did they perhaps refer only to conversational speech, of which the isolated utterances examined by Repp were not representative?

A careful reading of some of the source texts suggests that phoneticians did not intend to deny completely any acoustic manifestations of the articulatory release of the first stop. For instance, Abercrombie (1967) points out that, in /pt/ sequences, "There may be 'a little faintish smack' as the lips separate, as Abraham Tucker pointed out in 1773, but for practical purposes the stop is incomplete auditorily, and may be more specifically referred to as unexploded" (p. 146, our emphasis). And Jones (1956) points out that, in /pt/ and /bd/ sequences, "The /p/ and /b/ do not have normal plosion, that is to say no ^{h} or $^{\text{ə}}$ is heard when the lips are separated" (p. 155, our emphasis). These statements suggest that the authors were aware that a release burst of the first stop may occur, but that it is substantially weaker (i.e., of lower amplitude and shorter duration) than that of a released utterance-final stop in English, or that of a similar stop produced by a French speaker (Abercrombie, 1967, p. 147).

Given that such release bursts do occur, how common are they in fluent speech? Does the likelihood of their occurrence depend on the particular sequence of places of articulation of the two stops? Are the bursts so weak as to be difficult to detect by ear? The present study provides some answers to these questions. The last question, especially, relates to the criterion that phoneticians might have employed in the past for judging a stop to be "unreleased": The criterion may have been either a perceptual one ("unreleased" then means "without audible release burst") or a comparative phonetic one (a release burst may be heard, but it is not as strong as the burst of a "released" stop in the same or in some other language). We intended to examine whether there is any perceptual basis for calling stops preceding a nonhomorganic stop "unreleased."

ACOUSTIC OBSERVATIONS

Method

With three voiceless and three voiced stops in English, there are 24 possible sequences of two stops with different places of articulation. Of these, only four (/bd/, /gd/, /pt/, and /kt/) occur in word-final position, primarily in the past tense forms of verbs. All 24 sequences are permissible in word-medial position across a syllable boundary, but only two (/pt/ and /kt/) occur with any frequency, primarily in words of Romance origin. However, by including some compound words, we were successful in finding two examples of each of the 24 sequences in word-medial position.

We constructed meaningful sentences, each containing two of the words to be measured, and the subjects read from a typed list of these sentences. The sentences are shown in Appendix 1 with the critical words underlined. As can be seen, all stop sequences were immediately preceded and followed by a vowel, with primary stress on the preceding vowel. (Note that we were not concerned here with two-stop sequences across a word boundary, although two stops crossing a morpheme boundary in words such as bootcamp may be considered a rather similar instance.)

Six native speakers of American English, three male and three female, were selected as subjects. They were not informed about the purpose of the experiment, but were asked to first study the sentences and then read them at a normal conversational speed. Their productions were recorded on magnetic tape using a Sennheiser MKH 415T microphone, placed approximately 8 inches from the subject's lips, and a Crown SX 822 tape recorder. The recordings were then digitized at 10 kHz using the Haskins Laboratories pulse code modulation system, and the waveforms were displayed on an oscilloscope. We zeroed in on the closure periods in the critical words to determine whether or not a release burst of the first stop was present. If present, such bursts appeared as distinct spikes of a few milliseconds duration, roughly in the center of the closure period. A typical example is shown in Figure 1a, with the closure and the release bursts for both stops indicated for the utterance scapegoat, produced by a female speaker (CG). In some cases, the release bursts were of very low amplitude, and two of the subjects produced a few tokens containing multiple or exaggerated bursts, but the token shown in Figure 1a is representative of the majority of utterances containing release bursts.

Results

The frequency of occurrence of a release burst for the first stop in word-medial sequences is shown in Table 1. The columns represent the six possible sequences of two different places of stop articulation, while the rows represent the individual subjects. The voicing feature of the stops has been ignored in this analysis, so that the percentage in each cell is based on eight words. Looking at the means in the right margin, we see that, overall, 58 percent of the words contained a release burst of the first stop, with the average percentages for individual speakers ranging from 46 to 81 percent. It is further evident from the means in the bottom row that release bursts were not equally common in all consonant combinations. The main determinant was

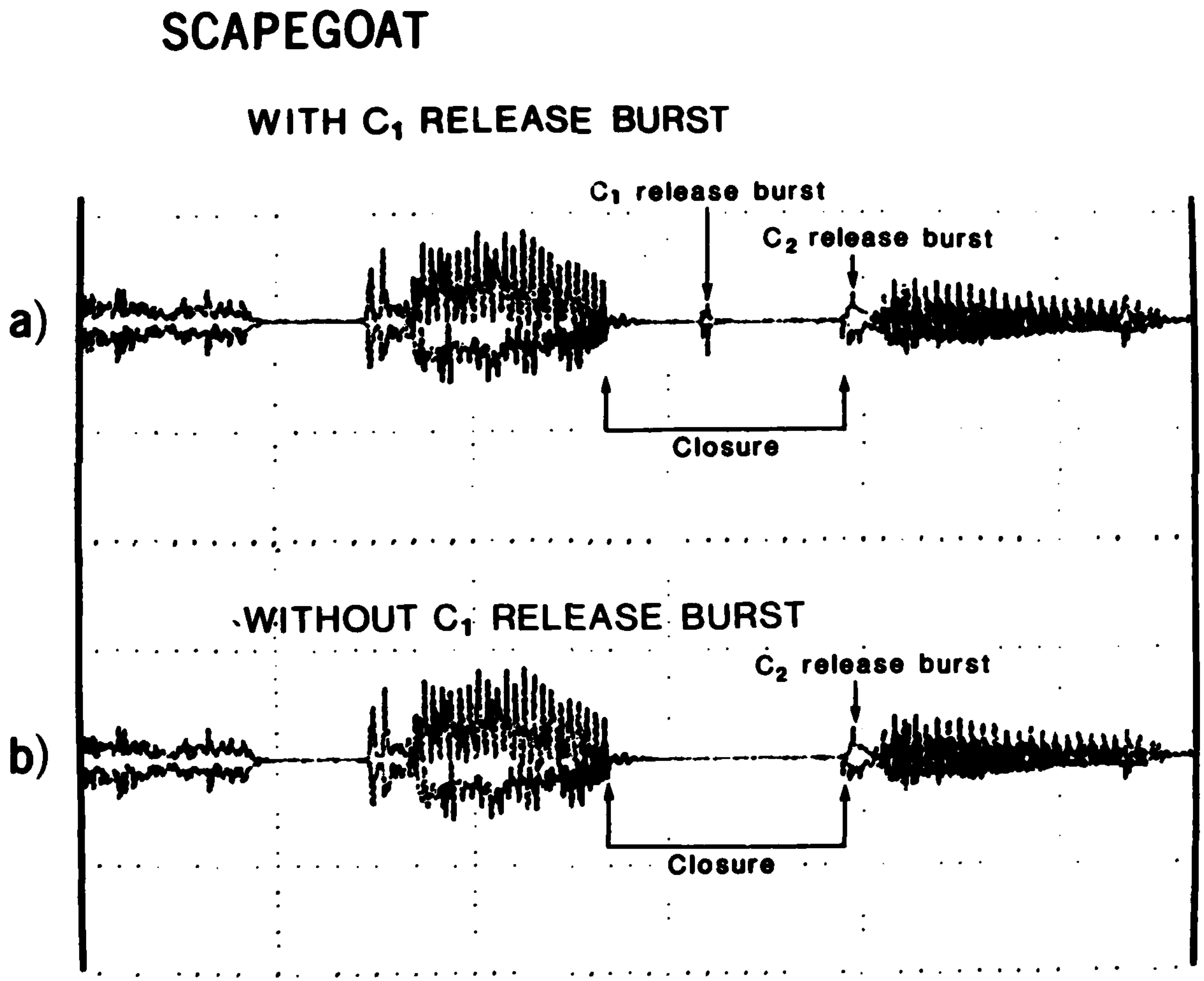


Figure 1. Oscillogram of the word <u>scapegoat</u> produced by a female speaker. The word is shown excised from its sentence context with the release burst of the first stop in place (above) and removed (below).

Table 1

Percentage of Words with C_1 Release Bursts

Place of Stop Articulation

	C_1:	ALV	VEL	VEL	LAB	ALV	LAB	Mean
	C_2:	LAB	LAB	ALV	ALV	VEL	VEL	
Speakers								
NM		25.0	25.0	50.0	12.5	75.0	87.5	45.8
AB		0.0	0.0	50.0	87.5	87.5	87.5	52.1
BR		0.0	0.0	37.5	87.5	87.5	100.0	52.1
CG		12.5	12.5	75.0	75.0	75.0	87.5	56.3
JM		0.0	25.0	87.5	87.5	87.5	75.0	60.4
RK		12.5	87.5	100.0	87.5	100.0	100.0	81.3
Mean		8.3	25.0	66.7	72.9	85.4	89.6	58.0

Table 2

Percentage of Words with C_1 Release Bursts

Place of Stop Articulation

	C_1:	Labial	Velar	Mean
	C_2:	Alveolar	Alveolar	
Speakers				
NM		100.0	75.0	87.5
AB		75.0	75.0	75.0
BR		100.0	100.0	100.0
CG		100.0	100.0	100.0
JM		100.0	25.0	62.5
RK		50.0	75.0	62.5
Mean		87.5	75.0	81.25

the place of articulation of the second stop. When the second stop was labial, release bursts of the first stop tended to be absent (except for one speaker's velar-labial sequences); when it was alveolar, release bursts were present in the majority of utterances; and when it was velar, release bursts were even more common. The place of articulation of the first stop seemed to play only a minor role, and we also observed that the voicing feature had no consistent influence on the occurrence of release bursts.[1]

Table 2 shows the same analysis for the word-final stop sequences (see Sentences 1-4 in the Appendix), with the columns representing the only two possible sequences of place of articulation, and the rows representing the same individual subjects. Again the voicing feature has been ignored so that the percentage in each cell is based on four words here, since no words containing stop sequences differing in voicing (e.g., /bt/, /kd/) occur in word-final position in English. The means in the right margin show that, overall, 81 percent of the words contained a release burst of the first stop, with the average percentages for individual speakers ranging from 63 to 100 percent. The means in the bottom row indicate that, as in word-medial position, the place of articulation of the first stop had no consistent effect.

Discussion

The pattern in these data can be understood by considering the articulatory maneuvers involved. When the second stop is labial, the speaker has the option of closing the lips before an earlier alveolar or velar closure is released, and if this option is followed, the release of the first stop occurs during the labial closure and therefore has minimal acoustic consequences. On the other hand, if the first stop is labial, although an alveolar or velar closure may be established before the lips are parted, the labial release, when it occurs, will generally produce a burst because there is no occlusion anterior to the lips. The occasional absence of a detectable burst may be due to changing local conditions (e.g., dryness of the lips) that affect sound generation. When one stop is alveolar and the other velar, we must take into account that the same articulator--i.e., the tongue--is involved. Even though, in principle, the tongue tip could establish contact with the palate before the tongue body releases its contact (and vice versa), this seems a difficult maneuver that speakers do not commonly employ. Our data show that release bursts occur both in alveolar-velar and velar-alveolar sequences, suggesting that the second closure is established shortly after the release of the first. If the closure periods of the two stops had overlapped, release bursts in velar-alveolar sequences should have been considerably less frequent because the velar release would have been silenced by the alveolar closure.

Our findings refute a strictly acoustic interpretation of the statement that the first stop in nonhomorganic two-stop sequences is unreleased. Contrary to that interpretation, which predicts the absence of release bursts, we have found that release bursts are generally present, at least in those two-stop sequences that occur most frequently in English (i.e., those in which the second stop is alveolar). There is no reason to believe that our results would not generalize to the more common case of two-stop sequences across a word boundary. In fact, a word boundary might be expected to increase the probability of occurrence of a release burst of the first stop. As for two-stop sequences in word-final position, which are typically cited in discussions of "unreleased" stops, our data show that release bursts of the first stop are actually more frequent than in word-medial position.

Although some authors (Abercrombie, 1967; Jones, 1956) mentioned faint release bursts, it is our impression that their occurrence has not been generally acknowledged. One reason for this may be that they are difficult to detect by ear. We conducted a brief experiment to address this issue.

BURST DETECTION EXPERIMENT

Method

Five typical utterances were chosen from speaker CG's productions, all containing release bursts of the first stop (cactus, ribcage, Edgar, bodkin, scapegoat). Using the Haskins Laboratories pulse code modulation system, we excerpted the words from their sentence context and then created a second version of each in which the release burst of the first stop was replaced with silence. Figure 1b shows this modified version of the word scapegoat, the original of which is displayed in Figure 1a.

We then constructed two discrimination tests. In the Yes/No test, each of the ten stimuli occurred ten times in random order, with interstimulus intervals (ISIs) of 3 sec. In the 2IFC test (two-interval forced-choice test), the two versions of each word were arranged in pairs, with the modified version either first or second. The resulting ten pairs occurred ten times in random order, with ISIs of 500 msec within pairs and 2 sec between pairs.

Nine subjects participated; they were the two authors and seven colleagues at Haskins Laboratories with varying amounts of phonetic training and experience. In the Yes/No discrimination test, they were provided with a written copy of the randomized tokens and were asked to indicate whether each stimulus did or did not contain a release burst of the first consonant in the two-stop sequence. In the subsequent 2IFC test, the subjects were asked to listen to each pair of words and then indicate which member, the first or the second, contained the release burst. The subjects were told that the bursts might be difficult to hear and listened to some examples before starting each test.

Results

The mean percentages of correct responses are shown in Table 3, with the five words displayed separately for the two tasks. We see that overall performance was poor on both tests, though better than chance (50 percent). The average score on the 2IFC test was only slightly higher than that on the Yes/No test, even though the 2IFC test, which was administered last, had the potential benefit of practice during the Yes/No test. Individual stimuli varied in difficulty, with cactus being near chance level while scapegoat (cf. Figure 1) reached a respectable 80 percent correct in the 2IFC task. We have not investigated in detail the acoustic properties that account for this variation, but two factors that are likely to play a role are the amplitude of the release burst and its temporal separation from the much stronger release burst of the second stop.

Table 3

Mean Percentage Correct Discrimination

	Discrimination Task	
Stimuli	Yes/No	2IFC
Cactus	49.4	55.0
Ribcage	61.1	58.3
Edgar	62.2	63.9
Bodkin	64.4	62.2
Scapegoat	67.8	80.5
Mean	61.0	64.0

There was also considerable variability between subjects. In the Yes/No test, the two authors performed at 83 and 85 percent correct, respectively, whereas the scores of the other seven listeners ranged from 45 to 66 percent correct. In the 2IFC task, the corresponding values were 89 and 79 for the authors and 50-67 for the other subjects. Thus, if one excludes the two subjects who had pre-experimental experience with the stimuli and perhaps knew better what to listen for, there is little evidence that even phonetically trained listeners can detect the faint release bursts of so-called "unreleased" stops. This is, then, the likely reason why the bursts were not noticed by some earlier authors who relied on their auditory impressions.

CONCLUSIONS

In this paper, we have reported some data relevant to the statement that, in English, stops followed by a different stop are "unreleased." We have examined several possible interpretations of that statement: (1) If it is interpreted as referring to articulation, it is clearly false. (2) If it is interpreted as referring to the acoustic signal, it is not generally true unless the definition of what is to count as a "release burst" is restricted to acoustic events of a certain minimal duration and amplitude. While such a restrictive definition may have been implicit in some previous discussions of "unreleased" stops, it should be noted that, on the contrary, the term "burst" is appropriately applied only to the signal portion excluded by such a definition--viz., to the brief transient generated by the stop release, exclusive of any following aspiration (cf. Dorman, Studdert-Kennedy, & Raphael, 1977; Fant, 1973). (3) If the statement is interpreted as referring to perception, it appears to be accurate in so far as stops preceding another stop in conversational speech have release bursts that are difficult to detect by ear. In this sense, the stops in this study were indeed "unreleased." (4) The possibility remains that some phoneticians have used the term "unreleased" in a purely contrastive sense. In this usage, even a stop with a detectable release burst might qualify as "unreleased" relative to some standard for "released" stops. The stops recorded by Repp (1980, in press), whose release bursts were from 10-40 msec long and quite detectable, may fall in this category. An obvious problem here is the absence of any clearly defined criterion separating the two classes.

These considerations illustrate the confusion that can result from terminology that is not only vague about the level of description to which it refers (Repp, 1981), but also insufficiently defined at the level intended. Many phonetic distinctions that are couched in acoustic terminology have been drawn at some remove from the speech signal. In that respect, the term "unreleased" is similar to the term "unaspirated," which is commonly applied to consonants, such as English [g], that exhibit a good deal of aspiration in the acoustic signal. While these terms may be sufficient for the field phonetician, they do not reflect the level of detail that acoustic phoneticians are concerned with, and therefore are of limited use.

We propose the following, more detailed classification, in which "release" is reinstated as an articulatory term:

(1) Unreleased: The occlusion is maintained, as in a stop preceding a homorganic stop or in many utterance-final stops with delayed release.
(2) Silently released: No release burst in the acoustic record.
(3) Inaudibly released: Visible release burst in records of the signal, but not readily detectable by ear.
(4) Weakly released: Release burst detectable by ear but clearly weaker than in (5).
(5) Strongly released: Release burst is followed by substantial aspiration or voicing.

In this scheme, successive classes are separated by different criteria: (1) and (2) by an articulatory criterion, (2) and (3) by an acoustic criterion, (3) and (4) by a perceptual criterion, and (4) and (5) by a criterion of phonetic contrast or classification.

In summary, our studies indicate that, in English, stops preceding a nonhomorganic stop in conversational speech are generally released inaudibly or silently, silent releases being particularly common when the following stop is labial. The observations of Repp (1980, in press), on the other hand, suggest that similar stops produced in isolated disyllables are typically weakly released.

REFERENCES

Abercrombie, D. Elements of general phonetics. Edinburgh: Edinburgh University Press, 1967.

Catford, J. C. Fundamental problems in phonetics. Bloomington: Indiana University Press, 1977.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. Perception & Psychophysics, 1977, 22, 109-122.

Fant, G. Speech sounds and features. Cambridge, Mass.: M.I.T. Press, 1973, 110-142.

Jones, D. An outline of English phonetics (8th ed.). Cambridge: W. Heffer and Sons, 1956.

Kenyon, J. S. American pronunciation. Ann Arbor: George Wahr Pub. Co., 1951.

Ladefoged, P. A course in phonetics. New York: Harcourt, Brace, Janovich Inc., 1975.

MacKay, I. R. A. Introducing practical phonetics. Boston: Little, Brown Pub. Co., 1978.

Repp, B. H. Perception and production of two-stop-consonant sequences. Haskins Laboratories Status Report on Speech Research, 1980, SR-63/64, 177-194.

Repp, B. H. On levels of description in speech research. Journal of the Acoustical Society of America, 1981, 69, 1462-1464.

Repp, B. H. Perceptual assessment of coarticulation in two-stop sequences. Haskins Laboratories Status Report on Speech Research, in press, SR-69.

FOOTNOTE

[1] We considered the possibility that the absence of release bursts in some tokens was due to the substitution of glottal stops for alveolar (and, perhaps, velar) stops. In the informal judgment of the first author, 22 utterances may have contained glottal stops. In 18 of these, the putative glottal stop preceded a labial stop. Release bursts were observed in 4 of these 18 tokens (22 percent), which is slightly higher than the overall incidence of 17 percent in this context (cf. Table 1). Thus, to the extent that glottal stops did occur, they did not change the pattern of our results.

APPENDIX 1

1. The old lady was mugged and robbed of her purse.

2. The fifth act of the play was a particularly hard one to direct in the small theater.

3. The little girl sobbed incessantly although her mother hugged and kissed her.

4. Peter is very apt on occasions to forget that the journals must be kept in the reference room.

5. The breakdown of his car made it very difficult for Edgar to get on and off campus for his classes.

6. Last night we relaxed with a bowl of popcorn and watched the movie "Dogday Afternoon" starring Al Paccino.

7. Even the Russians admit that straight vodka is real rotgut.

8. Considering he wasn't wearing a seatbelt, the driver was lucky to escape with only bruising to his ribcage and abdomen.

9. Nancy found a wonderful recipe for sugar-free cupcakes in her new cookbook.

10. The doctors did their rounds of the sickbay at 10 o'clock every morning.

11. The cactus was left so long in the small pot that it became completely rootbound and eventually died.

12. Bonnie and Clyde each carried a shotgun and left a bloodbath behind them after every bank robbery.

13. Dan bought a new puptent for his backpacking trip on the Appalachian Trail.

14. Gary's favorite sport is rugby but his talents make him a good football player too.

15. King Edward decided to abdicate and leave a respectable lagtime before his marriage to Mrs. Simpson.

16. Crockpots are excellent for cooking Chinese soups such as wonton, subgum, and eggdrop.

17. The oddball in our dorm was Egbert who permanently locked himself in his room so he wouldn't have to socialize with any of us.

18. The army has developed subterranean landing pads which allow helicopters to escape from the radar detection of invading aircraft.

19. The incessant burrowing of the new-born pigs had turned their pigpen into a huge mudpuddle.

20. One of Deborah's favorite hobbies is tapdancing, especially to jazz and ragtime music.

21. Some people claim that Nixon was only a scapegoat in the cover-up of C.I.A. scheming and subterfuge.

22. Margaret caught her 9 year old son trying to shoot a magpie with his popgun.

23. In the Fall, the catkins hanging outside the backdoor of the cottage were really beautiful.

24. My grandmother always inserted a hatpin or a bodkin into her cakes to see if they were ready to be removed from the oven.

25. The marine biologists made a movie about the development of tadpoles into frogs through a trapdoor mechanism on the side of the artificial pond.

26. The uniforms for the Governor and Subgovernor of India during the early 1900's could only be differentiated by the headgear and the collar markings.

27. It seemed that the menu for bootcamp consisted largely of potpies and oatmeal.

28. David tried to prevent the bogdown of his car by putting sacks under the wheels, but after a few attempts at moving it, it sank up to the hubcaps in the mud.

OBSTRUENT PRODUCTION BY HEARING-IMPAIRED SPEAKERS: INTERARTICULATOR TIMING AND ACOUSTICS*

Nancy S. McGarr+ and Anders Löfqvist++

Abstract. This study examined the organization of laryngeal control and interarticulator timing in the production of obstruents and obstruent clusters by three severely-profoundly deaf adults. Laryngeal activity was monitored by transillumination; temporal patterns of oral articulation (lips and tongue-palate) were recorded using an electrical transconductance technique. For each of the deaf speakers, an inappropriate laryngeal abduction gesture was often found between words, a pattern never observed for hearing speakers. At the same time, the deaf speakers differed from each other with respect to type of errors, variability, and interarticulator coordination. For the most intelligible speaker, the timing of glottal opening with respect to oral articulation was most like that observed for normals. The second deaf speaker often failed to observe voicing contrasts with respect to glottal opening. This subject was nevertheless consistent in producing most plosives without a glottal opening, and all fricatives with an opening gesture. For the third deaf speaker, the pattern of errors was more complex and included both missing and inappropriate glottal opening gestures.

INTRODUCTION

Production of voiceless obstruents requires intricate coordination of several articulatory systems. At the laryngeal level, an abduction/adduction gesture normally occurs to stop glottal vibrations and assist in the buildup of oral pressure. Supralaryngeal adjustments are also necessary to produce a closure or constriction. Thus, laryngeal and supralaryngeal articulations involve simultaneous activities that must be temporally coordinated. Differences in the relative timing of the laryngeal and oral gestures are used

*Parts of this paper were presented at the 101st Meeting of the Acoustical Society of America, Ottawa, May 18-22, 1981.

+Also Center for Research in Speech and Hearing Sciences, Graduate School and University Center, The City University of New York.

++Also Department of Phonetics, Lund University, Sweden.

Acknowledgment. We are grateful to Thomas Baer and Katherine S. Harris for comments on an earlier version of this paper, and to Kiyoshi Honda, David Zeichner, and Richard Sharkany for technical assistance during the experiments. This work was supported by NINCDS Grants NS-13617 and NS-13870, and by NIH Biomedical Research Support Grant RR-05596 to Haskins Laboratories.

in a wide variety of languages to produce contrasts in voicing and aspiration (cf. Lisker & Abramson, 1964; Löfqvist & Yoshioka, 1981).

Since the larynx is placed in an inaccessible and invisible position, it is reasonable to assume that coordination of interarticulator gestures is learned by auditory monitoring of the acoustic signal. Developmental studies suggest that children master sound contrasts requiring laryngeal adjustments (e.g., voicing and aspiration) by attending to their acoustic and perceptual consequences (Kewley-Port & Preston, 1974; Zlatin & Koenigsknecht, 1976; Gilbert, 1977; Macken & Barton, 1980). These studies also show that obstruent contrasts emerge relatively late in children's speech and that production is more variable in children than in adults. The acoustic cues for obstruents are complex, spread over time, and involve differences in the sound source and the spectral composition of the signal. For example, in the production of a voiceless fricative in a vocalic environment, the sound source changes from periodic to aperiodic and back to periodic. Similarly, a voiceless aspirated stop in the same environment is associated with the following sequence of source changes: periodic voicing during the preceding vowel, silence during the closure, transient noise, aspiration noise, periodic voicing during the vowel. In addition to being spread out over time, the acoustic attributes of obstruents often involve short-term spectral changes, where high frequency components play an important role. Examples of such attributes are release bursts and formant transitions for stop consonants, and spectra and transitions for fricatives.

Given the complex articulatory and acoustic/perceptual nature of voiceless obstruents, one would expect hearing-impaired speakers to have particular problems with this class of sounds. This is indeed the case, as shown by several descriptive and acoustic studies. For example, hearing-impaired speakers frequently fail to make the voiced-voiceless distinction (Hudgins & Numbers, 1942). In some studies, this substitution is reported as occurring to the voiced member of the pair (Heider, Heider, & Sykes, 1941; Carr, 1953; Millin, 1971; Smith, 1975), and at other times to the voiceless cognate (Mangan, 1961; Nober, 1967; Markides, 1970). At the acoustic level, several studies report a lack of voice onset time distinction for deaf speakers (Monsen, 1976; Mahshie, 1980), and also that closure or constriction duration is different from normals (Calvert, 1961; Osberger & Levitt, 1979). Production of obstruent clusters is also particularly difficult for hearing-impaired speakers (Hudgins & Numbers, 1942; Brannon, 1964; Smith, 1975). Reported error patterns for these blends include the dropping of one or more components of the cluster, or the adding of an adventitious segment, usually the neutral [ə], between the elements. Errors in voicing as well as in cluster production have been shown to affect the overall intelligibility of deaf speech deleteriously (Hudgins & Numbers, 1942).

While we may presume that it is some failure of interarticulator coordination that leads to these perceptual and acoustic results, it is not possible to infer the precise nature of the laryngeal and supralaryngeal events from the acoustic record. The purpose of the present investigation, therefore, was to make detailed observations on glottal and upper articulatory events during obstruent and obstruent cluster production in a comparison of hearing-impaired and normal speakers.

Table 1

Summary of Listener Judgments of the Productions by the Deaf Speakers. The Percentage of Correct Productions, the Percentage of Errors, and the Error Categories Are Shown.

	Deaf Speaker 1		Deaf Speaker 2		Deaf Speaker 3	
Single Obstruents	% Corr.	% Err.	% Corr.	% Err.	% Corr.	% Err.
peal	100		17	83 (b)	83	17 (b)
beak	100		100		33	67 (p)
teal	100		100		100	
deal	83	17 (t)		100 (t)		100 (t)
seal	100		100		100	
zeal	17	50 (dʒ) 33 (d)		100 (s)	67	33 (s)
paper	100			100 (b)		100 (b)
paper	100			100 (b)		100 (b)
chicken	50	33 (s) 17 (t)		100 (s)		67 (ʃ) 17 (z) 17 (ʒ)
jester	50	33 (ʃ) 17 (d)		100 (s)		67 (ʒ) 17 (t) 17 (ʃ)
Obstruent Clusters						
steal	83	17 (s)	83	17 (s)		83 (t) 17 (s)
jester	17	50 (ht) 33 (t)	17	33 (t) 33 (h) 17 (ht)		100 (ʃ)
less tea	100		33	67 (sd)	100	
messtent	100		100		100	

METHOD

Subjects

The subjects were three congenitally and profoundly deaf adults[1] (mean pure tone average 90 dB+ ISO in the better ear) and a hearing subject who served as a control. All of the hearing-impaired subjects had received at least part of their training in oral schools for the deaf. None had any additional handicaps besides deafness.

Speech samples from each of the hearing-impaired speakers were analyzed in two ways: First, two listeners (the authors) made broad phonetic transcriptions of the test words produced by the deaf subjects. Of particular interest was the voicing status of the obstruents and obstruent clusters. In general, the listeners agreed in their judgments, and the results are summarized in Table 1. Second, speech samples were rated for overall intelligibility by a listener highly experienced with the deaf. Following the format of the rating scale for intelligibility (Subtelny, 1975), deaf speaker 1 could be characterized as intelligible with the exception of a few words or phrases. The speech of deaf speaker 2 could be characterized as difficult to understand although the gist of the content could be understood. Deaf speaker 3 was difficult to understand with only isolated words or phrases intelligible.

Linguistic Material

The linguistic material is presented in Table 2. It consisted of voiced and voiceless obstruents, and also of obstruent clusters. Since the transillumination technique requires a free passage for the light from the fiberscope to the glottis, front vowels and labial and dental consonants were chosen. The subjects read the material from randomized lists in the carrier phrase "Say ... again." Six repetitions of each token were obtained.

Table 2

The Linguistic Material

peal	paper	chicken	steal
beak		jester	less tea
teal			messtent
deal			
seal			
zeal			

Procedure

Laryngeal activity was monitored by transillumination (Sonesson, 1960). A flexible fiberscope inserted through the nose and held in position by a headband provided illumination of the larynx. The amount of light passing through the glottis was sensed by a phototransistor placed on the surface of the neck just below the cricoid cartilage and coupled to the skin by a light-tight enclosure. The transillumination signal was recorded on one channel of a multichannel instrumentation tape recorder. During the recording session, the view of the larynx was monitored through the fiberscope in order to detect movements and fogging of the lens.

Temporal information on laryngeal articulatory movements obtained by transillumination has been shown to be practically identical to similar information obtained by fiberoptic filming of the larynx (Yoshioka, Löfqvist, & Hirose, 1981; Löfqvist & Yoshioka, 1980). Transillumination is thus an excellent tool for studying laryngeal behavior in speech. It has a better temporal resolution than fiberoptic filming and video recording. Data collection and processing are quick and easy, and larger amounts of material can be handled than with any other method available for laryngeal investigations.

Temporal patterns of oral articulation were recorded using an electrical transconductance technique (cf. Karlsson & Nord, 1970). The electrodes of a modified laryngograph were placed on the upper and lower lips, respectively. Onset and offset of lip or tongue-palate contact could then be identified from changes in the electrical signal. The signal was recorded on another channel of the instrumentation recorder.

Conventional acoustic recordings were obtained simultaneously using a direction-sensitive microphone. The voice signal was recorded in direct mode on the tape recorder. As described above, the recordings of all productions by the hearing-impaired speakers were later used to obtain listener judgments.

For each token, a number of measurements were made relevant to motor control. Measurements of closure and constriction duration were made from the transconductance signal representing labial or tongue-palate contact. Implosion was defined as the onset of labial/tongue-palate contact; release as the offset of contact. Defining these events in physiological terms is particularly helpful in the case of deaf speakers, since closure or constriction duration may be very difficult to measure in the acoustic waveform.

For laryngeal articulation, the occurrence of peak glottal opening served as the reference point. This point marks the end of the abduction and the beginning of the adduction of the vocal folds, and is under motor control. EMG recordings from internal laryngeal muscles have indicated a reciprocal pattern of activation for the posterior cricoarytenoid and the interarytenoid muscles in the control of glottal opening in single voiceless obstruents (cf. Hirose, 1976; Hirose, Yoshioka, & Niimi, 1978; Löfqvist & Yoshioka, 1980). This justifies the use of peak glottal opening as a reference point in studies of laryngeal articulation in speech.

Measures of interarticulator timing were defined in two ways. First, the interval from onset of labial or tongue-palate contact to peak glottal opening was calculated. This measurement provides an estimate of the relationship between onset of constriction or closure and the beginning of the adduction of the vocal folds. It is useful since it highlights differences in timing between obstruents, e.g., stops and fricatives (Löfqvist & Yoshioka, 1981). A second measurement of interarticulator timing was the interval from peak glottal opening to offset of labial or tongue-palate contact. This measure shows the relationship between onset of glottal adduction and release, and is particularly useful in examining timing differences between different stop categories (Löfqvist, 1980). The physiological measurements were supplemented by acoustic measurements of voice onset time for stops. All measurements were made interactively on a computer.

RESULTS

Single Obstruents

Figure 1 shows representative tokens of the hearing subject's productions of voiceless and voiced stops. A glottal abduction/adduction gesture is seen in the transillumination signal for the voiceless stop but not for the voiced cognate. Patterns of interarticulator timing are noted in the relationship between events recorded in the signals representing labial/tongue-palate contact and glottal opening, respectively. For the voiceless plosive, peak glottal opening occurs at the oral release, indicated by the offset of lip contact and the release burst. This pattern is the same as that found for other speakers of American English (Löfqvist & Yoshioka, 1981).

Figure 2 shows selected tokens of the same utterances produced by deaf speaker 2. Several patterns are different from normal. First, closure duration is considerably longer for the deaf than the hearing speaker's productions. Second, there is evidence of an inappropriate glottal gesture. The deaf speaker made a glottal abduction/adduction gesture immediately preceding the test word, before the onset of lip closure for the initial stop. Thus, for both productions, glottal adduction starts before lip closure, and the glottis is in a position suitable for voicing at the release of the oral closure. The abduction/adduction gesture between words was fairly typical of the other deaf speakers as well, but was never observed for the hearing speaker.

From these raw data, a number of measurements were made that are summarized in Figures 3-4 and also in Figures 6-9. Line 1 in these figures shows the mean duration of closure of constriction. Line 2 shows, as a histogram, the number of instances of a glottal opening associated with the obstruent production. The third row shows the first measure of interarticulator timing--the interval between implosion and peak glottal opening. The second measure of interarticulator timing is the interval between peak glottal opening to release, indicated in numerals below the third row. A negative value implies that peak glottal opening occurred after the release. The presentation follows our general impression in rank order of overall speaker intelligibility: (1) the hearing speaker; (2) deaf speaker 1 (felt to be the most intelligible deaf speaker); (3) deaf speaker 2, and (4) deaf speaker 3.

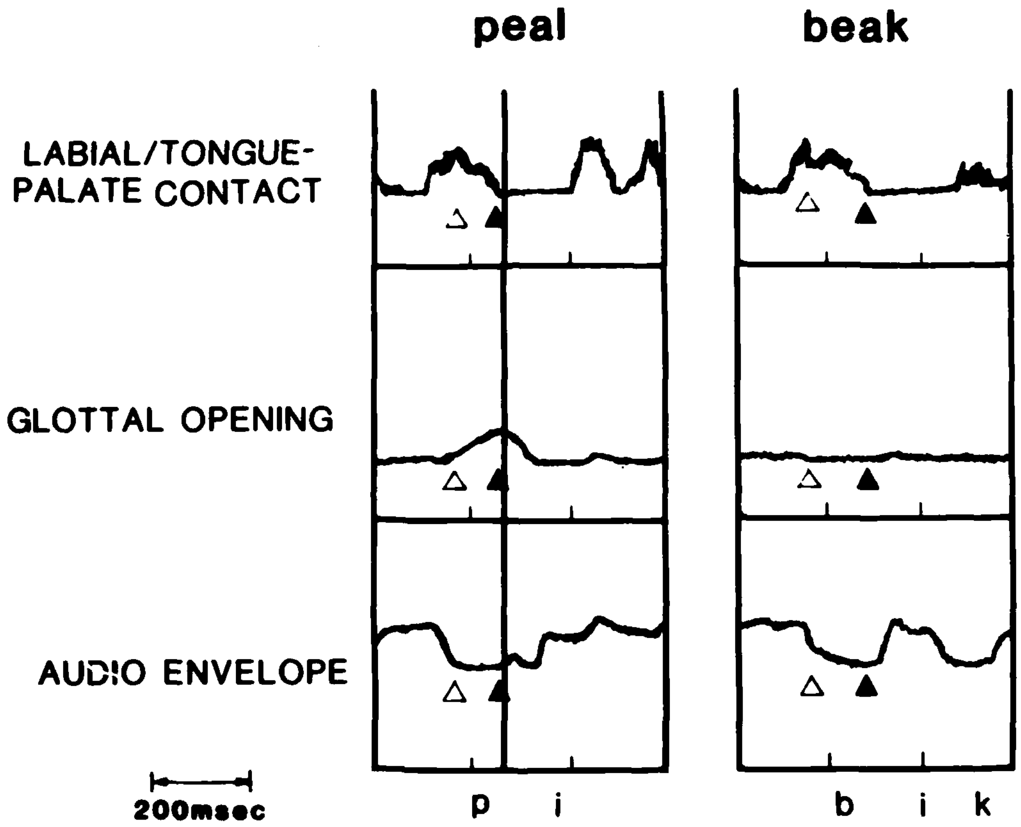


Figure 1. Records of the hearing speaker's productions of the utterances "peal" (left), and "beak" (right). Curves represent labial/tongue-palate contact (top), glottal opening (middle), and audio envelope (bottom). Onset of labial closure for the word initial labial stops in "peal" and "beak" is marked by △ , release of oral closure by ▲ . The vertical line indicates the time at which peak glottal opening occurs.

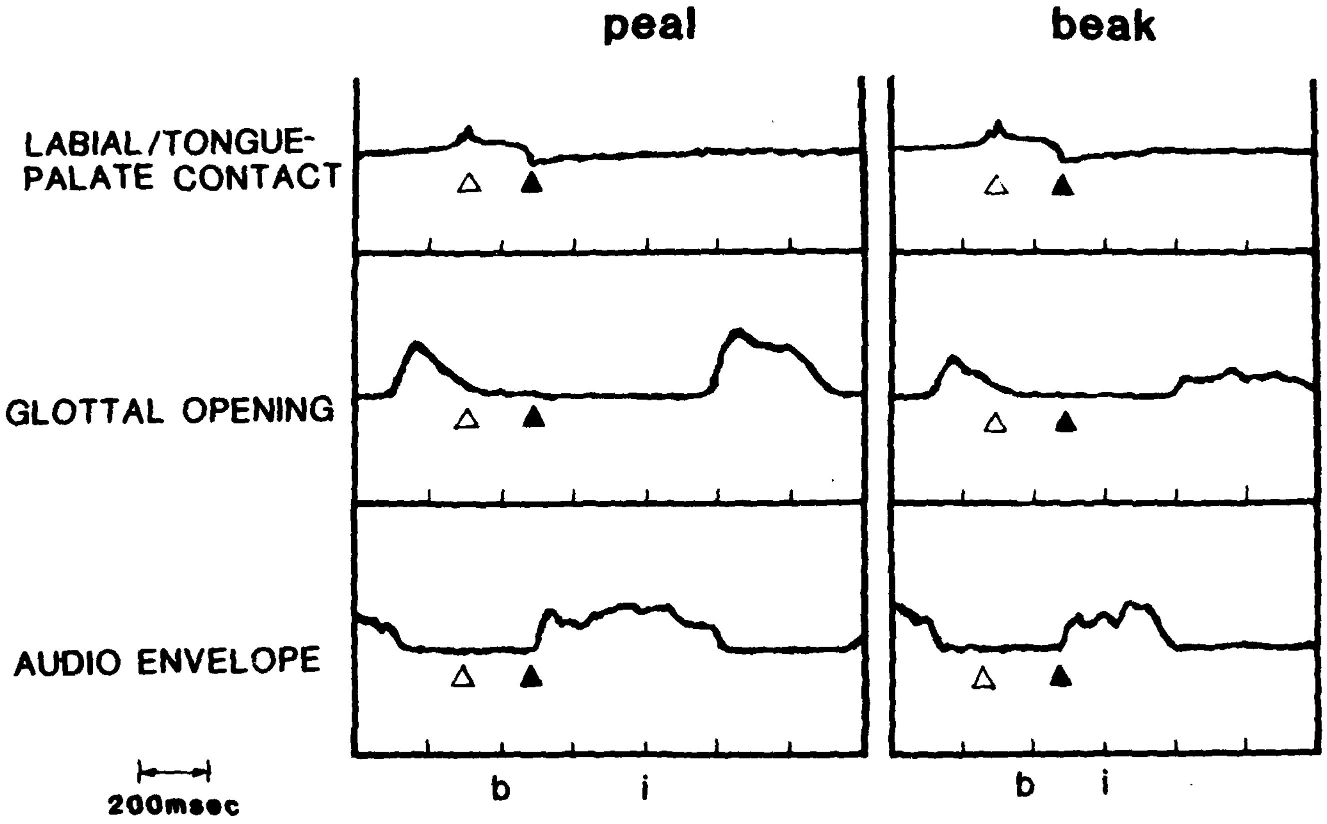
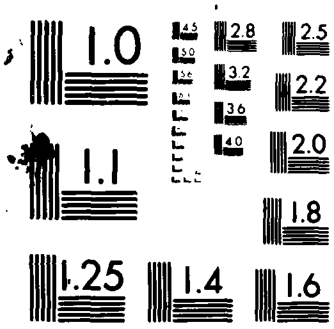


Figure 2. Records of deaf speaker 2's productions of the utterances "peal" (left), and "beak" (right). Symbols as in Figure 1.

AD-A111 385   HASKINS LABS INC NEW HAVEN CT   F/G 6/16
STATUS REPORT ON SPEECH RESEARCH, JULY-DECEMBER 1981.(U)
1981   PHS-HD-01994
UNCLASSIFIED   SR-67/68(1981)   NL

2 3



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

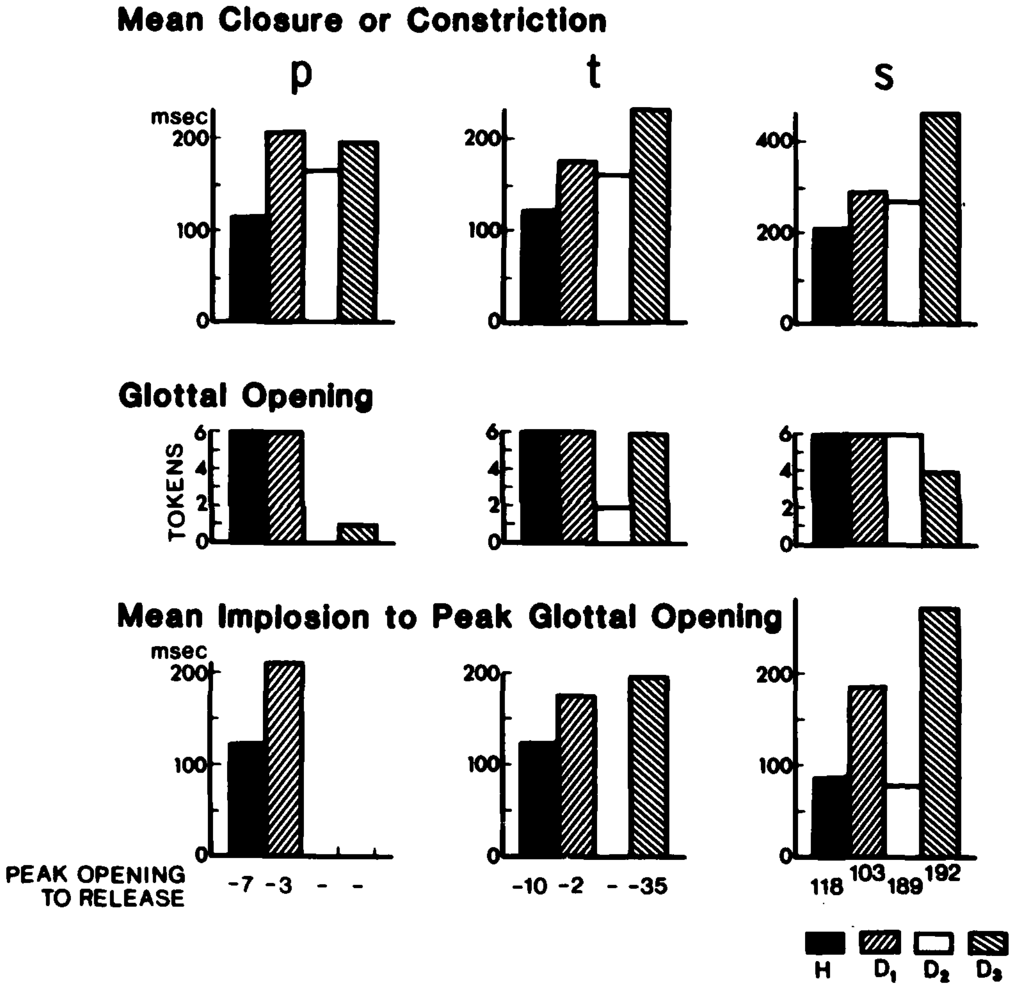


Figure 3. Summary of measurements for single voiceless obstruents. See text for further details on measurements.

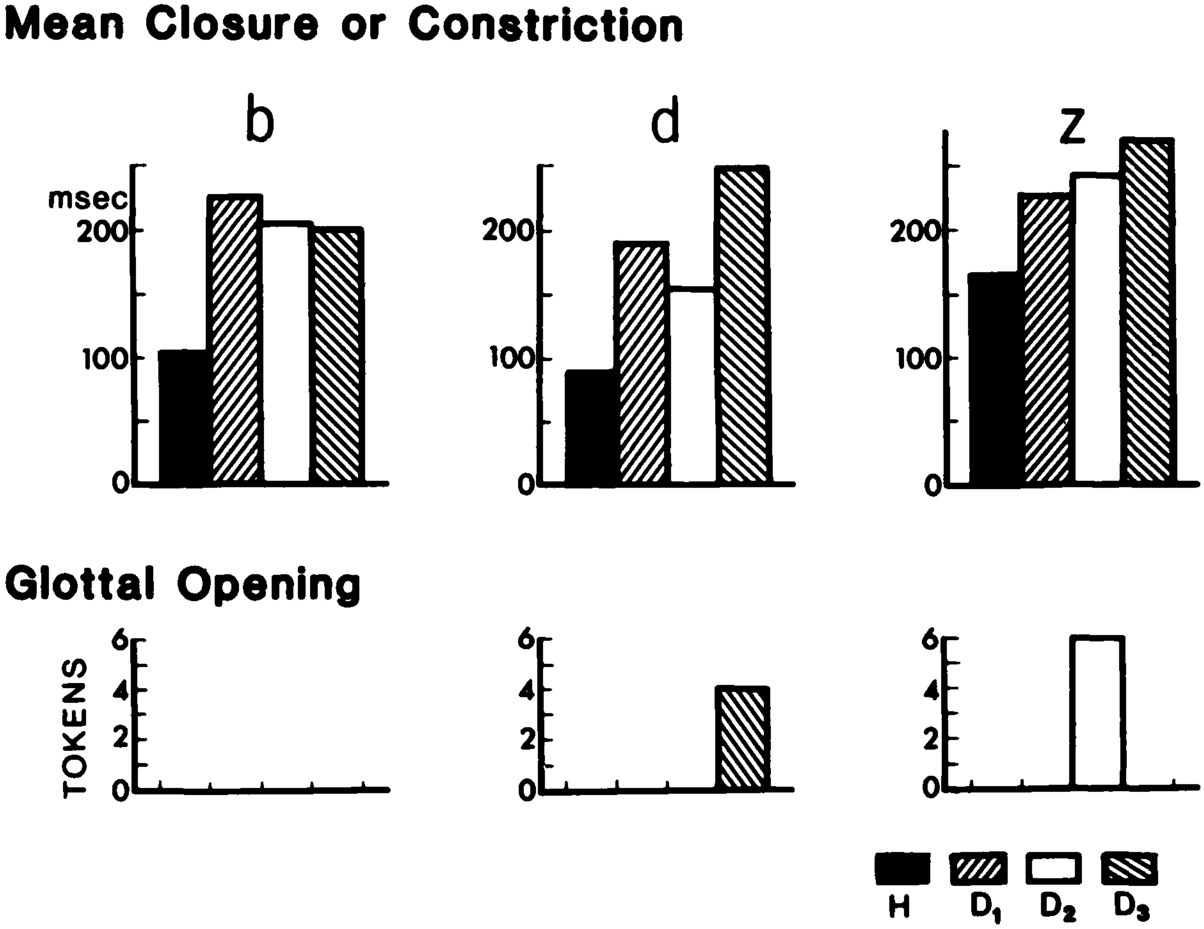


Figure 4. Summary of measurements for single voiced obstruents.

Results for the single voiceless and voiced obstruents are summarized in Figures 3 and 4, respectively. Closure or constriction duration was always longer for the deaf subjects than for the hearing subject, consistent with previous reports. As is typical for hearing speakers, closure or constriction duration was longer for voiceless than for voiced segments. For the deaf speakers, the duration measurements for the voiceless and voiced segments overlapped (see also below, Figure 5).

The number of tokens for which a glottal gesture occurred are shown in line 2. These gestures were always correct for the hearing speaker and deaf speaker 1. That is, for single voiceless obstruents, each token was characterized by a single abduction/adduction gesture; for single voiced obstruents, there was no laryngeal gesture. For the other deaf speakers, the pattern varied. Deaf speakers 2 and 3 used an appropriate laryngeal gesture more often for the alveolar than for the bilabial obstruents. We will discuss the voiced obstruents of these speakers below.

With respect to interarticulator timing, both the hearing speaker and deaf speaker 1 showed nearly similar patterns for all segments. For voiceless stops, the interval from implosion to peak glottal opening tends to be similar to closure duration. This means that peak glottal opening and oral release almost coincide. Thus, these two speakers both show a small negative number for the second measure of interarticulator timing, i.e., the interval from peak glottal opening to release. Even though the closure durations for the deaf speaker are prolonged overall, the relative timing of oral and laryngeal gestures is indistinguishable from normal. For the voiceless fricatives of these two speakers, the interval from implosion to peak glottal opening is roughly half of the duration of the oral constriction. Peak glottal opening thus occurs about 100 msec before release.

Deaf speaker 2 was inconsistent in production, since in most cases there was no active glottal opening gesture for the stops. For the fricative, there was an appropriate laryngeal gesture and interarticulator timing was more normal. For deaf speaker 3, we again find an inconsistent pattern. For the labials, there was no glottal opening, whereas for the alveolars, a glottal opening gesture was made. The interarticulator timing in these cases is similar to normal. For /t/, the glottis did not begin to close until about 35 msec after the oral release, which is somewhat long, although not totally unusual. For the fricative, although the durations [illegible] long overall, the relative timing pattern was similar to the pattern obtained for normal speakers.

Usually one does not discuss laryngeal-oral coordination for voiced obstruent production, but since deaf speakers are known to produce voiceless for voiced segments, we have also examined these productions. Figure 4 shows these data. Here, we again find evidence that deaf speakers may use an inappropriate laryngeal abduction gesture for the production of some voiced sounds, but as before, the speakers are inconsistent in this aberrant pattern.

When the deaf speakers produced the appropriate laryngeal gestures for voiceless stops, their overall pattern of interarticulator timing resembled that of normals. Specifically, the oral release and peak glottal opening tend to correspond in time. For fricatives, peak glottal opening precedes offset

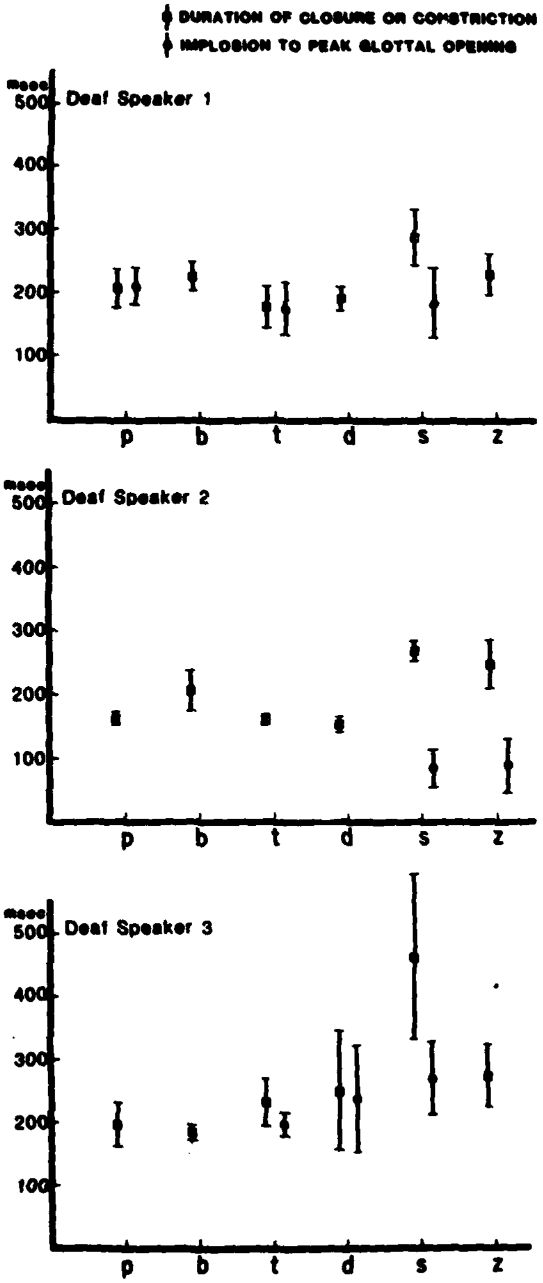


Figure 5. Plot of articulatory measurements for the three deaf speakers' productions of single obstruents. Means and standard deviations are shown.

of tongue-palate contact as has been observed for normals. But a rather unexpected finding was obtained for these deaf subjects. In general, the laryngeal gesture for the voiceless fricative /s/ was produced correctly more often than for the voiceless plosives. For example, as shown in Figures 3 and 4, deaf speaker 2 consistently contrasted stops and fricatives at the glottal level--the former were nearly always produced with a closed glottis, while for the latter, the glottis was always open. However, as shown in Figure 5, the deaf speakers were unlike the normal in that they were highly variable in their production from token to token. Standard deviations for the deaf speakers were, in many cases, fairly large. For the hearing speaker, the standard deviations were quite small--on the order of 10-25 msec, and therefore not included in the figure.

For all test words described above, obstruents were produced in the word-initial position. An allophonic variation in American English is that voiceless stops following a stressed vowel are unaspirated. Therefore, we also examined stops produced in two different positions of a bisyllabic word--"paper," where p_1 is stressed and p_2 is unstressed. These data are shown in Figure 6. The timing pattern for the inital stops in this test word was essentially the same as that described above for all speakers' production of a single voiceless stop. For p_2, the pattern is similar for the hearing subject and deaf speakers 2 and 3. Closure duration was shorter in these cases and there was a tendency not to use an abduction gesture in production. However, deaf speaker 1 produced both initial and medial stops in an almost identical way, with aspiration in both cases.

Table 3

Measurements of Voice Onset Time for Single Stop Consonants (msec, n=6)

		H	D_1	D_2	D_3
p	$\bar{x}$	84	87	8	29
	s	8.1	5.6	3.9	6.9
b	$\bar{x}$	15	16	11	25
	s	3.5	4.0	3.3	6.8
t	$\bar{x}$	121	83	20	69
	s	13.8	16.1	3.0	6.3
d	$\bar{x}$	23	47	21	59
	s	4.3	19.6	3.5	6.7
p_1	$\bar{x}$	68	77	11	25
	s	7.4	4.5	4.8	6.8
p_2	$\bar{x}$	14	71	20	
	s	6.7	3.4	3.8	

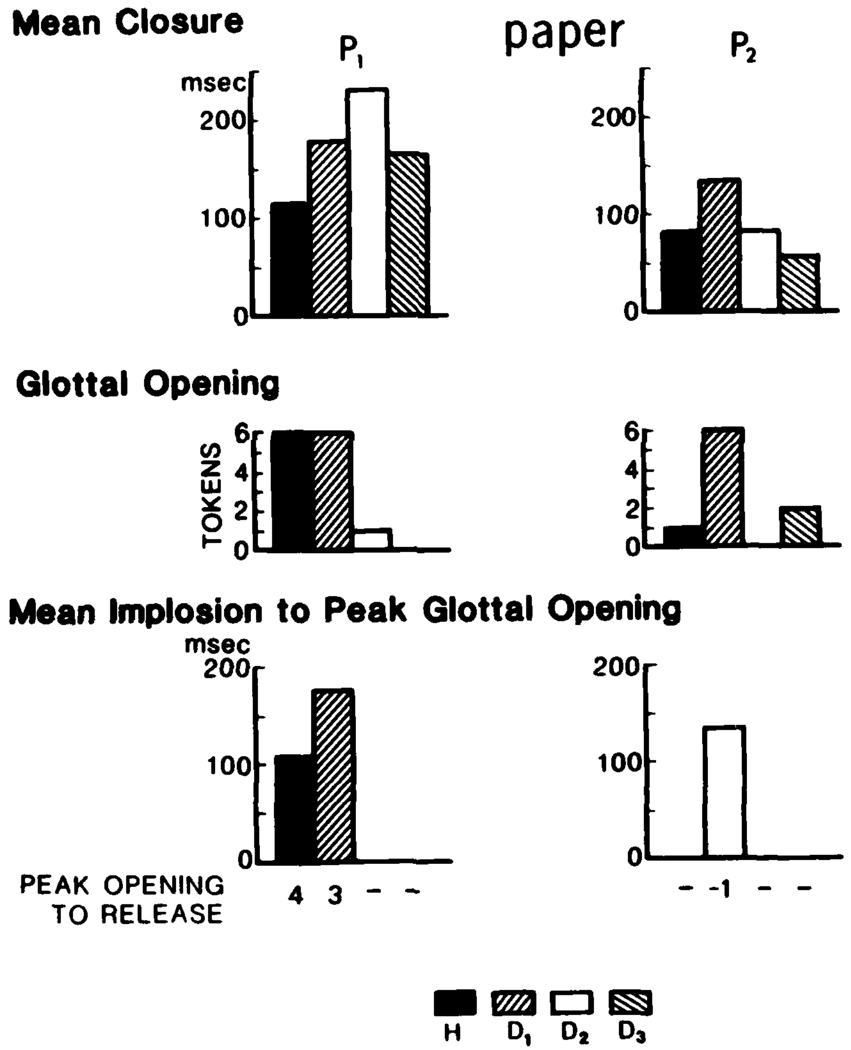


Figure 6. Summary of measurements for the two stops in "paper."

Table 3 shows measurements of voice onset time for single stops. These acoustical measurements match fairly well with the physiological data, i.e., voice onset time was generally longer when a glottal gesture was found. However, in contrast to the physiological data, the standard deviations for the acoustic measurements were fairly small.

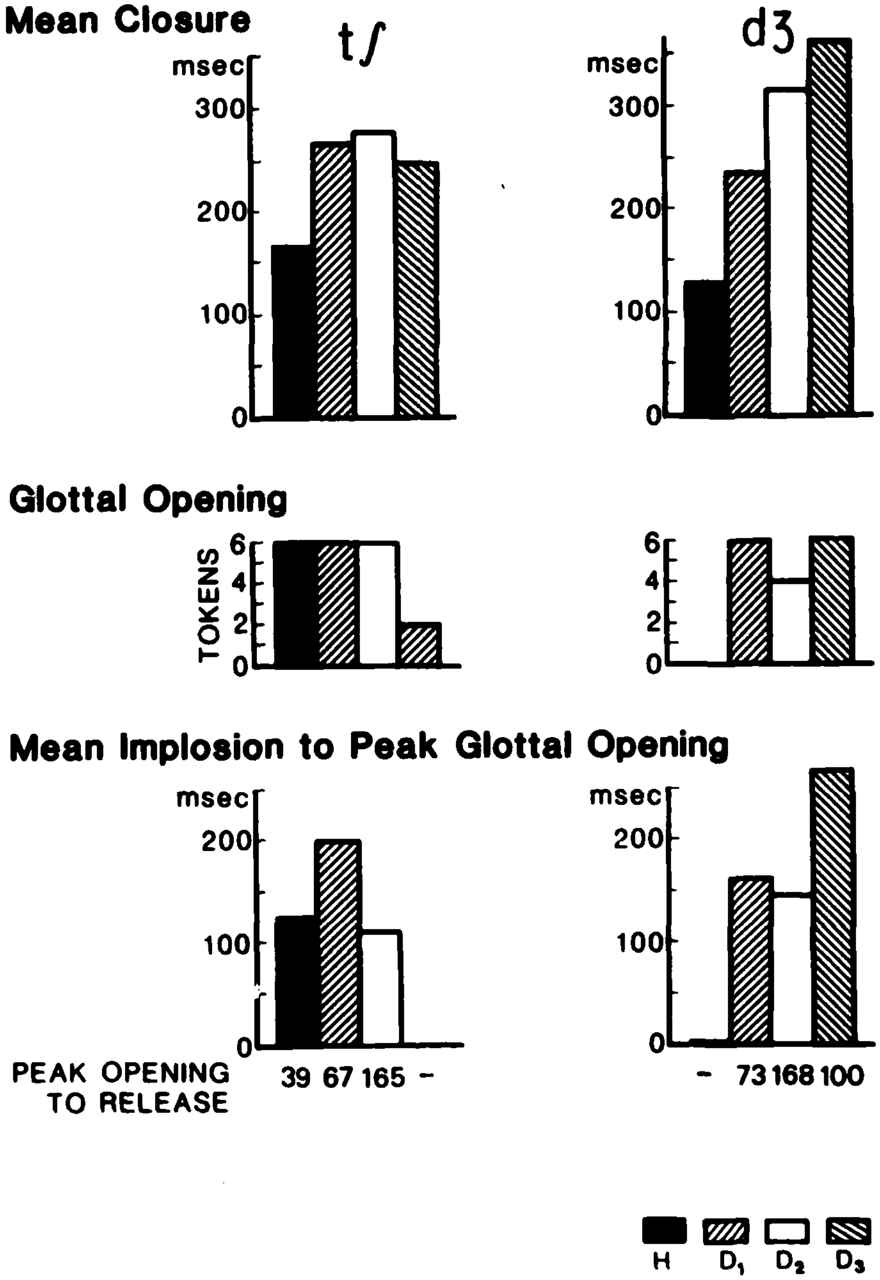
Data for affricates are shown in Figure 7. These segments are known to be particularly difficult for deaf speakers to produce. For the hearing subject, the stop closure and the fricative portion of the voiceless affricate were 39 and 126 msec, respectively, with peak glottal opening occurring during the fricative portion. In contrast, for the deaf speakers there was in most cases no stop component. Consequently, the timing pattern resembled that of a fricative. All deaf speakers produced the voiced affricates with a laryngeal abduction gesture.

Clusters

Clusters have not been studied much in the speech of the hearing impaired. The common /st/ cluster was examined in the word inital position and in the medial unstressed position of a two-syllable word. Figure 8 shows only one component of the cluster since we were often unable to identify two separate gestures for the hearing-impaired speakers. Consequently, these productions mostly resemble patterns described above for the single voiceless fricatives. For the hearing speaker, when a voiceless unaspirated stop followed a fricative, peak glottal opening is timed during the fricative segment and the glottis begins to close before the stop component begins. Deaf speaker 1 tended to use a timing pattern for an aspirated stop with peak glottal opening at release. In some cases, two opening gestures occurred--one for the fricative and one for the stop. For deaf speakers 2 and 3, in most cases, interarticulator timing for the word initial cluster more closely resembled that observed for single fricatives. These timing patterns were similar to normal in that peak glottal opening occurred during the fricative portion. No clear pattern emerges for these speakers' productions of /st/ in "jester."

We finally turn to clusters with either a word or morpheme boundary within the cluster, see Figure 9. In the first case, that of the word boundary ("less tea"), we would expect that the word inital stop /t/ would be aspirated since aspiration here is a way of signaling that a word boundary occurs between the /s/ and the /t/. In fact, all of the speakers, with the exception of deaf speaker 2, produced these tokens with two separate glottal gestures--one for the fricative and one for the stop. The patterns of deaf speaker 2 are consistent with the previous observation that this deaf speaker produced most stops without glottal opening, although for these test words, he nevertheless respected the word boundary. The pattern of interarticulator timing is similar to that observed for other tokens of fricatives and aspirated stops.

Turning now to the effect of the morpheme boundary, the pattern for the fricative segment is similar to that for other single fricatives. For the stop segment, only the hearing speaker shows evidence of a separate laryngeal adjustment. Deaf speakers 1 and 2 did not use a glottal opening. For deaf speaker 3, no stop segment could be identified.



Mean Closure

Glottal Opening

Mean Implosion to Peak Glottal Opening

Figure 7. Summary of measurements for the affricates.

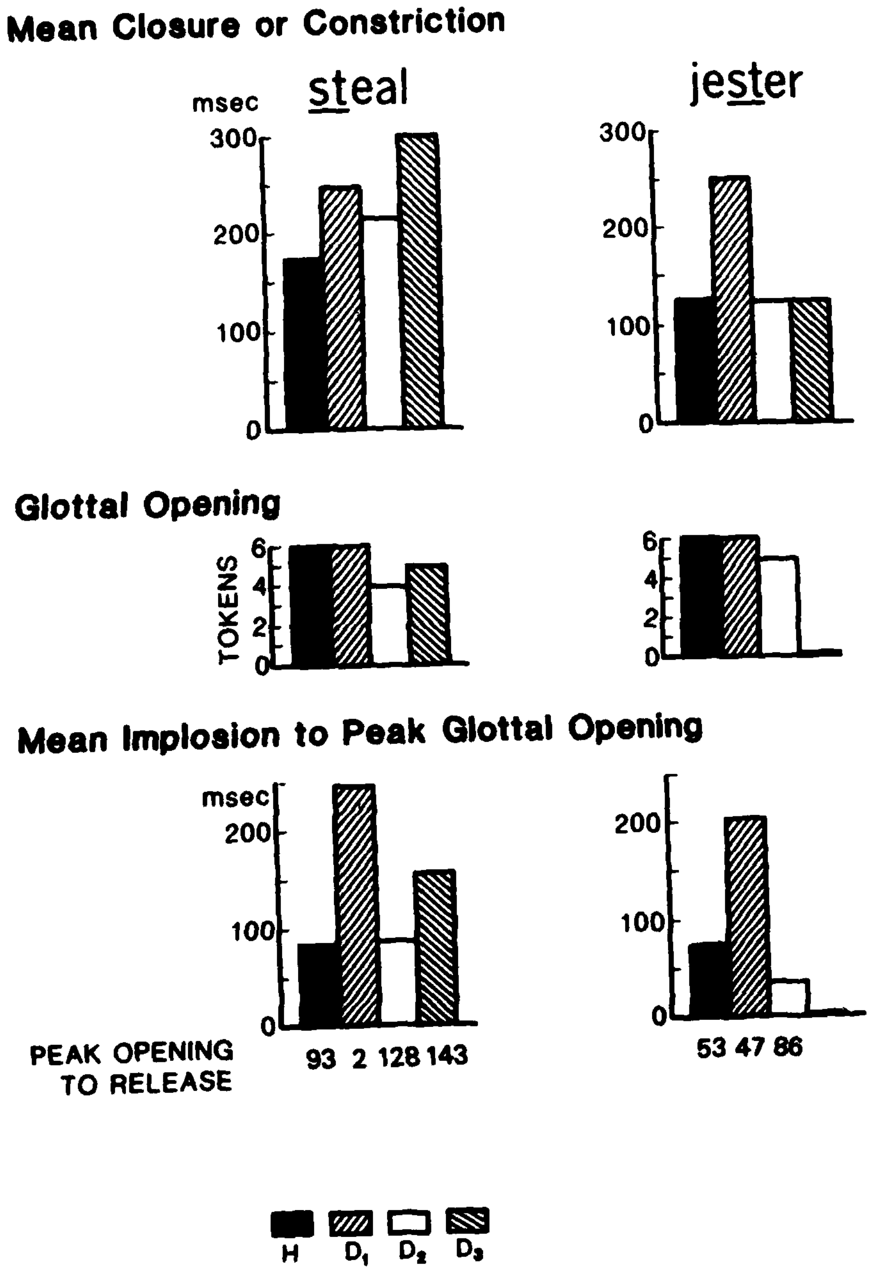


Figure 8. Summary of measurements for /st/ clusters.

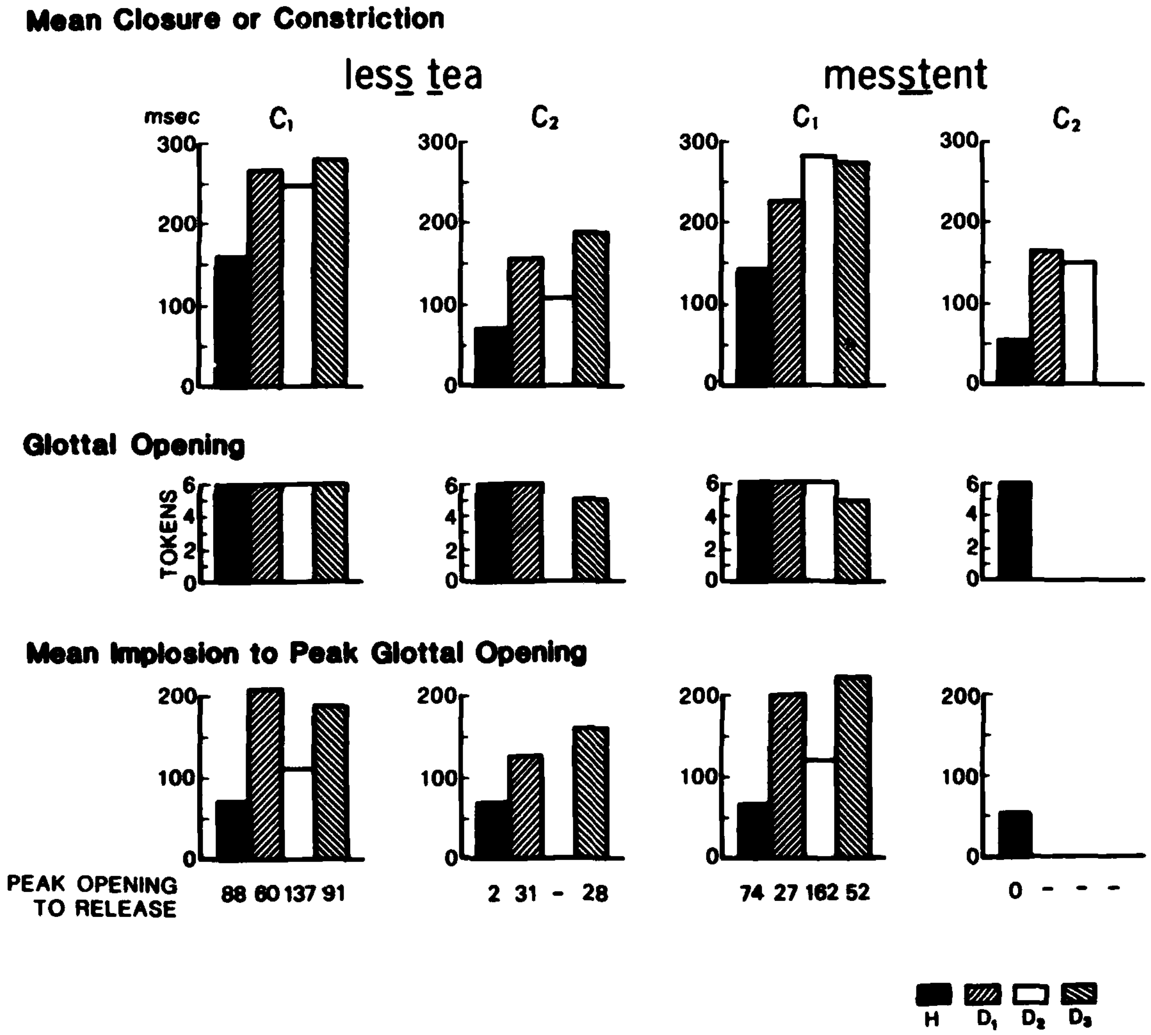


Figure 9. Summary of measurements for clusters with word and morpheme boundary within the cluster.

DISCUSSION

Normal speakers consistently use different patterns of laryngeal-oral coordination for voiceless stops and fricatives (Löfqvist & Yoshioka, 1981). Onset of glottal abduction generally tends to coincide with onset of oral closure or constriction, unless preaspiration occurs, in which case glottal abduction precedes implosion. For aspirated stops, peak glottal opening occurs at the release of the oral closure. This ensures a delay in voice onset time and also allows a high rate of air flow for generation of frication noise immediately after the release. In fricatives, the peak glottal opening occurs closer to the onset of the oral constriction. The velocity of the abduction gesture is higher for fricatives than for stops and the size of the glottal opening also tends to be larger for the fricatives. These differences in laryngeal control and interarticulator timing are most likely related to different aerodynamic requirments at implosion and release for fricatives and aspirated stops, respectively. The hearing speaker in this study followed these patterns.

The deaf subjects showed both similarities and dissimilarities with respect to normal speakers. The most obvious dissimilarity was failure to produce the voiced-voiceless distinction. The deaf speakers either made a glottal gesture when none was required or omitted the glottal gesture. Furthermore, even when a laryngeal gesture was produced, its timing relative to oral articulatory events could be more or less like normal. This pattern varied considerably among deaf speakers.

Not surprisingly, deaf speaker 1, the most intelligible, closely followed the normal pattern. For aspirated stops, peak glottal opening consistently occurred at the oral release. The same strategy was used in production of the second stop in the word "paper," although in this case, the phonological rules of American English dictate that aspiration is not necessary. On the other hand, while the timing for single fricatives was often produced correctly, the /st/ clusters showed different patterns of interarticulator timing. One example of this occurrence is illustrated by the /st/ cluster in "steal" where relative timing was observed to be like that for an aspirated stop. Again, this speaker uses an aspirated stop inappropriately--in this example as part of a segment cluster.

Deaf speaker 2 differs from normal in still a grosser fashion. Stops were consistently produced without laryngeal activity while fricatives were usually produced with an appropriate glottal gesture. For these latter cases, the interarticulator timing was relatively correct. Turning to deaf speaker 3, we note both incorrect and highly variable productions. However, when the relative timing is preserved between the articulators, the absolute duration of articulatory events is longer than those found for hearing speakers. This pattern of increased duration has often been noted in the speech of the deaf (Hudgins & Numbers, 1942; Calvert, 1961; Osberger & Levitt, 1979). In relation to these findings, it is interesting to note that hearing speakers, when deprived of auditory feedback, also show evidence of increasing duration (Borden, 1980).

Another characteristic that marks the speech of the deaf as different from normal is variability in production at the physiological level. This variability appears to be an important factor in the speech of the deaf suggesting that deaf speakers, even the less intelligible, do not produce an utterance in quite the same way each time it is perceived to be in error. However, we also observed that even when speakers were judged to be correct in their productions, there was considerable variability from token to token. These results are consistent with electromyographic data obtained for oral articulatory timing (tongue - lips) of a deaf talker (McGarr & Harris, in press). Variability in production was noted less at the acoustic level (VOT measurements), although fairly large standard deviations for deaf speaker productions have been reported (Monsen, 1976). Such inconsistencies in production may be one reason why listeners find the speech of the deaf so difficult to understand.

As mentioned above, all deaf speakers were more successful in producing fricatives than stops. These results differ from those reported in the literature (Nober, 1967; Smith, 1975; Levitt, Stromberg, Smith, & Gold, 1980). On the one hand, we find our results perplexing since one would expect that fricatives, because of their high frequency spectra and articulatory invisibility, would be difficult for severely-profoundly deaf speakers to perceive and thus to produce. Alternatively, on the physiological level, one might postulate that voiceless fricatives, for example, require less precise interarticulator timing than voiceless stops. At the laryngeal level, the deaf speaker need only open the glottis, even if in a fairly stereotypic way as demonstrated by our subjects, and then direct the air stream in an outward direction. The distortion of the /s/ in the speech of the hearing impaired may thus more accurately reflect poor placement of the upper articulators rather than inappropriate laryngeal adjustments. Indeed, it is well known that normally the /s/ is produced at the level of the upper articulators with both channel and wake turbulence, the former being generated by the grooved portion of the tongue, and the latter generated when the airstream strikes the teeth. Deaf speakers are known to have difficulty positioning the tongue for correct place of articulation (Huntington, Harris, & Sholes, 1968; McGarr & Harris, in press). Plosives, on the other hand, demand particularly fine interarticulator coordination between the larynx and the upper articulators and more precise management of the air stream.

The operation of the larynx in speech is analogous to that of an air valve, whereby the valve must be opened for voiceless sounds to let some air escape, and must also be closed at the appropriate times in order to preserve the breath-stream. Studies of the respiratory patterns of deaf speakers have shown that these subjects evidence at least two kinds of problems. The first is that they initiate phonation at too low a level of vital capacity, and, also that they produce a reduced number of syllables per breath (Forner & Hixon, 1977; Whitehead, in press). A second problem is mismanagement of the volume of air by inappropriate valving at the laryngeal level. Laryngeal valving has two functions: articulatory and phonatory. For the former, aerodynamic studies of deaf speech production do not consistently show that hearing-impaired speakers produce obstruents with abnormally high air flow rates (Whitehead, in press). One might infer phonatory valving problems from some descriptive studies that often ascribe breathy voice quality to deaf speakers (Hudgins & Numbers, 1942; Monsen, Engebretson, & Vemula, 1978; Stevens, Nickerson, & Rollins, in press). The results of the present study suggest valving problems of a somewhat different nature. That is, during pauses between words, each of the deaf speakers in this study inappropriately opened the glottis. Whether they actually took a breath, as is suggested in the early work of Hudgins (1937), or simply wasted air cannot be ascertained directly from our data. However, we would argue that the latter is more likely since the glottal abduction gesture was smaller and shorter in duration between words than between utterances. This pattern differs from one hypothesized by Stevens et al. (in press). Based on spectrographic analysis of deaf children's productions, these authors proposed that the glottis is closed during pauses between words.

Turning to acoustics and perception, we find a rather straightforward relationship between physiological records and acoustic measurements for stops. The relationship between the physiological measurements and the listener judgments was not always direct. Perception of both voiced and voiceless obstruents could be found for tokens with and without a correct laryngeal gesture. For example, for the productions of deaf speaker 2, listeners heard /b/ for /p/, the common voiced for voiceless substitution, when no glottal opening was found, cf. Table 1 and Figure 3. However, for the alveolar stops of the same speaker, listeners reported a voiceless sound in all cases, including those without a glottal abduction. From Table 3 it appears that VOT was only 20 msec for these stops.

These results are not too surprising, since a straightforward relationship between physiology and listener judgments is unlikely in such a complex phenomenon as the voiced/voiceless distinction. This mismatch between physiological records and listener judgments of deaf speakers has also been noted by Mahshie (1980). Although in controlled studies using synthetic speech, VOT has been shown to be an important determiner for the voiced/voiceless distinction, in real speech there are a host of acoustic cues that may be co-responsible for this perception. Measurements along one single acoustic dimension cannot be readily expected to predict listener responses when other acoustic variables are not held constant, since interactions have repeatedly been shown to occur. Examples of such interactions that affect the perception of the voiced-voiceless distinction in stops are amplitude and duration of aspiration (Repp, 1979), and speech tempo and closure duration (Port, 1979; Fitch, 1981; see also Miller, 1981). Our VOT values for the deaf speakers were in the range of 20 - 30 msec, where interactions and boundary shifts are most likely to occur. This may be another reason why listeners to deaf speech have difficulty making judgments of particular phonetic segments.

Earlier, we argued that because the larynx is placed in an inacessible and invisible position, mastery of laryngeal articulation is arrived at by the acoustic signal. The deaf speakers in this study all sustained severe-profound hearing losses suggesting that oral-laryngeal articulation would be exceedingly difficult in light of reduced auditory acuity. In fact, deaf speakers are often said to place their articulators fairly accurately especially for those places of articulation that are highly visible, but fail to coordinate the movements between several articulators. Our data show that this notion of deaf speech is in part correct, yet our subjects were also capable of executing appropriate glottal gestures. We would argue that this is in part due to low frequency residual hearing that conveys some voicing information as well as tactile feedback.

There are other findings in studies of deaf speech that are also perplexing and not satisfactorily accounted for by either residual hearing or taction: prepausal lengthening (Reilly, 1979), and pitch declination (Breckenridge, Note 1). If auditory monitoring of one's own voice was the sole prerequisite for the establishment of these phenomena, one would not necessarily expect to find them in profoundly deaf speakers. Quite possibly, they may be due to intrinsic factors of the speech production system. This idea may also account for why interarticulator timing was sometimes correct for the hearing-impaired subjects of this study. Laryngeal articulatory movements overall are rather stereotypic and restricted to abduction and adduction. For example, production of a voiceless fricative involves opening the glottis and letting air through. This bears some resemblance to non-speech activities such as blowing and respiration. For the latter, it is reasonable to assume that there exist respiratory-laryngeal linkages whereby glottal abduction and adduction are automatically coordinated with respiratory activity. Speech production in both normals and the deaf most likely utilizes such linkages, although the details are unknown at present.

REFERENCE NOTE

1. Breckenridge, J. *Declination as a phonological process*. Unpublished manuscript, Bell Laboratories, 1977.

REFERENCES

Borden, G. J. Use of feedback in established and developing speech. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3). New York: Academic Press, 1980, 223-242.

Brannon, J. *Visual feedback of glossal motions and its influence upon the speech of deaf children*. Unpublished doctoral dissertation, Northwestern University, 1964.

Calvert, D. *Some acoustic characteristics of the speech of profoundly deaf individuals*. Unpublished doctoral dissertation, Stanford University, 1961.

Carr, J. An investigation of the spontaneous speech sounds of five-year old deaf-born children. *Journal of Speech and Hearing Disorders*, 1953, *18*, 22-29.

Fitch, H. Distinguishing temporal information for speaking rate from temporal information for intervocalic stop consonant voicing. *Haskins Laboratories Status Report on Speech Research*, 1981, *SR-65*, 1-32.

Forner, L., & Hixon, T. J. Respiratory kinematics in profoundly hearing-impaired speakers. *Journal of Speech Hearing and Research*, 1977, *20*, 373-408.

Gilbert, J. H. A voice onset time analysis of apical stop production in 3-year olds. *Journal of Child Language*, 1977, *4*, 103-110.

Heider, F., Heider, G., & Sykes, J. A study of the spontaneous vocalizations of fourteen deaf children. *Volta Review*, 1941, *43*, 10-14.

Hirose, H. Posterior cricoarytenoid as a speech muscle. *Annals of Otology, Rhinology and Laryngology*, 1976, *85*, 343-342.

Hirose, H., Yoshioka, H., & Niimi, S. A cross-language study of laryngeal adjustments in consonant production. *Annual Bulletin* (Research Institute of Logopedics and Phoniatrics, University of Tokyo), 1978, *12*, 61-71.

Hudgins, C. V. Voice production and breath control in the speech of the deaf. *American Annals of the Deaf*, 1937, *82*, 338-363.

Hudgins, C. V., & Numbers, F. C. An investigation of the intelligibility of the speech of the deaf. *Genetic Psychology Monographs*, 1942, *25*, 289-392.

Huntington, D., Harris, K. S., & Sholes, G. An electromyographic study of consonant articulation in hearing-impaired and normal speakers. *Journal of Speech and Hearing Research*, 1968, *11*, 147-158.

Karlsson, I., & Nord, L. A new method of recording occlusion applied to the study of Swedish stops. *STL-QPSR* 2/3, 1970, 8-18.

Kewley-Port, D., & Preston, M. Early apical stop production: A voice onset time analysis. *Journal of Phonetics*, 1974, *2*, 195-210.

Levitt, H., Stromberg, H., Smith, C., & Gold, T. The structure of segmental errors in the speech of deaf children. *Journal of Communication Disorders*, 1980, *13*, 419-441.

Lisker, L., & Abramson, A. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 1964, *20*, 384-422.

Löfqvist, A. Interarticulator programming in stop production. *Journal of Phonetics*, 1980, *8*, 475-490.

Löfqvist, A., & Yoshioka, H. Laryngeal activity in Swedish obstruent clusters. *Journal of the Acoustical Society of America*, 1980, *68*, 792-801.

Löfqvist, A., & Yoshioka, H. Interarticulator programming in obstruent production. *Phonetica*, 1981, *38*, 21-34.

Macken, M., & Barton, D. The acquisition of the voicing contrast in English: A study of voice onset time in word-initial stop consonants. *Journal of Chi'd Language*, 1980, *7*, 41-74.

Mahshie, J. *Laryngeal behavior in hearing impaired speakers*. Unpublished doctoral dissertation, Syracuse University, 1980.

Mangan, K. Speech improvement through articulation testing. *American Annals of the Deaf*, 1961, *106*, 391-396.

Markides, A. The speech of deaf and partially hearing children with special reference to factors affecting intelligibility. *British Journal of Disorders of Communication*, 1970, *5*, 126-140.

McGarr, N. S., & Harris, K. S. Articulatory control in a deaf speaker. In I. Hochberg, H. Levitt, & M. J. Osberger (Eds.), *Speech of the hearing impaired: Research, training, and personnel preparation*. Washington D.C.: A. G. Bell Association, in press.

Miller, J. L. The effect of speaking rate on segmental distinctions: Acoustic variation and perceptual compensation. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, N.J.: Erlbaum, 1981.

Millin, J. Therapy for reduction of continuous phonation in the hard-of-hearing population. *Journal of Speech and Hearing Disorders*, 1971, *36*, 496-498.

Monsen, R. The production of English stop consonants in the speech of deaf children. *Journal of Phonetics*, 1976, *4*, 29-42.

Monsen, R., Engebretson, A. M., & Vemula, N. Some effects of deafness on the generation of voice. *Journal of the Acoustical Society of America*, 1978, *66*, 1680-1690.

Nober, H. Articulation of the deaf. *Exceptional Children*, 1967, *33*, 611-621.

Osberger, M. J., & Levitt, H. The effect of timing errors on the intelligibility of deaf children's speech. *Journal of the Acoustical Society of America*, 1979, *66*, 1316-1324.

Port, R. The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics*, 1979, *7*, 45-56.

Reilly, A. P. *Syllabic nucleus duration in the speech of hearing and deaf children*. Unpublished doctoral dissertation, The City University of New York, 1979.

Repp, B. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. *Language and Speech*, 1979, *22*, 173-189.

Smith, C. Residual hearing and speech production of deaf children. *Journal of Speech and Hearing Research*, 1975, *18*, 795-811.

Sonesson, B. On the anatomy and vibratory pattern of the human vocal folds. *Acta Oto-laryngologica*, 1960, *Supplement 156*.

Stevens, K. N., Nickerson, R., & Rollins, A. Suprasegmental and postural aspects of speech production and their effect on articulatory skills and intelligibility. In I. Hochberg, H. Levitt, & M. J. Osberger (Eds.), *Speech of the hearing impaired: Research, training, and personnel preparation*. Washington, D.C.: A. G. Bell, in press.

Subtelny, J. Speech assessment of the deaf adult. *Journal of the Academy of Rehabilitative Audiology*, 1975, *8*, 110-116.

Whitehead, R. Some respiratory and aerodynamic patterns in the speech of the hearing impaired. In I. Hochberg, H. Levitt, & M. J. Osberger (Eds.), *Speech of the hearing impaired: Research, training, and personnel preparation*. Washington, D.C.: A. G. Bell, in press.

Yoshioka, H., Löfqvist, A., & Hirose, H. Laryngeal adjustments in the production of consonant clusters and geminates in American English. *Journal of the Acoustical Society of America*, 1981, *70*, 1615-1623.

Zlatin, M., & Koenigsknecht, R. *Development of the voicing contrast:* A comparison of voice onset time in perception and production. *Journal of Speech and Hearing Research*, 1976, *19*, 93-111.

FOOTNOTE

[1] For convenience in the following discussion, we will call the speech characteristics of the group "deaf speech" and the speakers of "deaf speech" will be called deaf." By making this identification, we acknowledge that not all persons who sustain severe to profound hearing losses produce this characteristic speech.

ON FINDING THAT SPEECH IS SPECIAL*

Alvin M. Liberman+

Abstract. A largely unsuccessful attempt to communicate phonologic segments by sounds other than speech led my colleagues and me to ask why speech does it so well. The answer came the more slowly because we were wedded to a "horizontal" view of language, seeing it as a biologically arbitrary assemblage of processes that are not themselves linguistic. Accordingly, we expected to find the answer in general processes of auditory perception to which the acoustic signal had been made to conform by appropriate regulation of the movements of articulation. What we found was the opposite: specialized processes of phonetic perception that had been made to conform to the acoustic consequences of the way articulatory movements are regulated. The distinctively linguistic function of these specializations is to provide for efficient perception of phonetic structures that can also be efficiently produced. To assume that a phonetic specialization exists accords well with a "vertical" view of language in which the underlying activities are seen as coherent and distinctive. Recent evidence for such special processes comes from experiments designed to investigate the integration of cues.

I welcome this opportunity to talk to my fellow psychologists about a subject that has, I think, been too much taken for granted. The subject is perception of phonetic segments, the consonants and vowels that lie near the surface of language. My aim is to promote the hypothesis that perception of those segments rests on specialized processes. These support a phonetic mode of perception, they serve a distinctively linguistic function, and they are part of the larger specialization for language.

*In press, American Psychologist.

+Also University of Connecticut and Yale University.

Acknowledgment. This paper is based on a Distinguished Scientific Contribution Award address given at the meeting of the American Psychological Association, Los Angeles, California on August 25, 1981. Preparation of the paper, and much of the research on which it is based, was supported by the National Institute of Child Health and Human Development (HD 01994) and by a Biomedical Research Support Grant (RR05596). At Haskins Laboratories it is hard to know what is owed and to whom. I would, however, especially acknowledge my debt to Franklin S. Cooper, with whom I have been closely associated for 35 years. For help with this paper I thank Louis Goldstein, Isabelle Liberman, Virginia Mann, Sharon Manuel, Ignatius Mattingly, Patrick Nye, Bruno Repp, and Michael Studdert-Kennedy. I am grateful to J. A. Fodor for making available to me an early draft of his monograph, "The Modularity of Mind," which I found particularly relevant and stimulating.

The phonetic specialization is apparently adapted to the singular code by which phonetic structure is connected to sound, a code that owes its character to the way the segments of the structure are articulated and coarticulated by the organs of the vocal tract. Not surprisingly, then, phonetic processes incorporate a link between perception and production. With that as key, an otherwise opaque code becomes perfectly transparent: diverse, continuous, and tangled sounds of speech are automatically perceived as a scant handful of discrete and variously ordered segments. Moreover, the segments are given in perception as distinctively phonetic objects, without the encumbering auditory baggage that would make them all but useless for their proper role as vehicles of language.

But we do take speech and its acoustic nature for granted, so much so that it is, I suspect, hard to see why perception of phonetic segments should require processes of an other-than-auditory sort, and even harder, perhaps, to imagine what it might mean to perceive those segments as phonetic objects, free of a weighty burden of auditory particulars. It may help, then, to begin by recounting my experience with an attempt to transmit phonologic information by purely auditory means. That experience exposed for me the problem that a phonetic specialization might solve, though it did not, of course, reveal how the solution is achieved, nor did it show that the solution requires specialized processes. Evidence bearing on those matters is reserved for later sections.

Perceiving Phonologic Segments in the Auditory Mode: An Assumption That Failed

In the mid-Forties I began, together with colleagues at Haskins Laboratories, to design a reading machine for the blind (Cooper, 1950; Nye, 1963; Studdert-Kennedy & Cooper, 1966). This was, or was to have been, a device that would scan print and use its contours to control an acoustic signal. At the outset we assumed that our machine had only to produce, for each letter, a pattern of sound that was distinctively different from the patterns for other letters. Blind users would presumably learn to associate the sounds with the letters and thus come, in time, to read. The rationale, largely unspoken, was an assumption about the nature of speech--to wit, that the sounds of speech represent the phonemes (roughly, the letters of the alphabet) in a straightforward way, one segment of sound for each phoneme. Accordingly, the perception of speech was thought to be no different from the perception of other sounds, except as there was, in speech, a learned association between perceived sound and the name of the corresponding phoneme. Why not expect, then, that arbitrary but distinctive sounds would serve as well as speech, provided only that the users had sufficient training?

Given that expectation, we were ill prepared for the disappointing performance of the nonspeech signals our early machines produced. So we persisted, seeking to increase the perceptual distinctiveness of the sound alphabet and also the ease with which its units would form into words and sentences. But our best efforts were unavailing. No matter how we patterned them, the sounds evoked a clutter of auditory detail that subjects could not readily organize and identify. This discouraged the subjects, but not me, for I had faith that the difficulty would ultimately yield to practice and the principles of learning. What loomed as a far more serious failing was that modest increases in rate caused the unit sounds to dissolve into an imperspicuous buzz. Indeed, this happened at rates barely one tenth those at which the discrete units of phonetic structure can be conveyed by speech.

Having come, thus, to the conclusion that we should try to learn from speech, we began to study it. But our hope at that early stage was only that we might find principles of auditory perception, hitherto unnoticed, that the language system had somehow managed to exploit.[1] These would not only be interesting in their own right, but also useful in enabling us to overcome the practical difficulty we had been having, since the auditory principles we hoped to find could presumably be applied to the design of nonspeech sounds our reading machine might be made to produce.

What I did not for a long time understand was that our practical difficulty lay, not in our having failed to find the right principles of auditory perception, but, much deeper, in our having failed to see that the principles we sought were simply not auditory. Perhaps I should have arrived at that understanding earlier had I not been in the grip of a misleading assumption that had decisively shaped my thinking about speech, language, and, indeed, almost anything else I might have found psychologically interesting. I was the more misled because the assumption reflected what I took to be the received view; in any case, I had never thought to question it.

In casting about for a word to characterize the view I speak of, I hit on "horizontal" as being particularly appropriate, only to discover that J. A. Fodor (Note 1) had chosen the same word to describe what I take to be much the same view. Apparently, we have here a metaphor whose time has come. As applied to language, the metaphor is intended to convey that the underlying processes are arranged in layers, none of them specific to language. On that horizontal orientation, language is accounted for by reference to whatever combination of processes it happens to engage. Hence our assumption, in the attempt to find a substitute for speech, that perception of phonologic segments is normally accomplished, presumably in the first layer, by processes of a generally auditory sort--that is, by processes no different from those that bring us the rustle of leaves in the wind or the rattle of a snake in the grass. To the extent we were concerned with the rest of language, we must have supposed, in like manner, that syntactic structures are managed by using the most general resources of cognition or intelligence. There were surely other processes on our minds when we thought about language--attention, memory, learning, for example--the exact number and variety depending on just which aspects of language activity our attention was directed to at the moment. But all the processes we might have invoked had in common that none was specialized for language. We were not prepared to give language a biology of its own, but only to treat it as an epiphenomenon, a biologically arbitrary assemblage of processes that were not themselves linguistic.

The opposite view--the one to which I now incline--is, by contrast, vertical. Seen this way, language does have its own biology. It is a coherent system, like echolocation in the bat, comprising distinctive processes adapted to a distinctive function. The distinctive processes are those that underlie the grammatical codes of syntax and phonology; their distinctive function is to overcome the limitations of communicating by agrammatic means. To appreciate those limitations, we need only consider how little we could say if, as in an agrammatic system, there were a straightforward relation between message and signal, one signal, however elaborately patterned, for each message. In such a system, the number of messages to be communicated could be no greater than the number of holistically and distinctively different signals that can be efficiently produced and perceived; and surely that number is very small, especially when the signal is acoustic. What the processes of syntax and phonology do for us, then, is to encode an unlimited number of messages into a very limited number of signals. In so doing, they match our message-generating capabilities to the restricted resources of our signal-producing vocal tracts and our signal-perceiving ears. As for the phonetic part of the phonologic domain, which is the subject of this paper, I will suggest that it, too, partakes of the distinctive function of grammatical codes, and that it is, accordingly, also special. (For further discussion, see Mattingly & Liberman, 1969; Liberman & Studdert-Kennedy, 1978; Liberman, 1970.)

The Special Function of the Phonetic Mode

To produce a large, indeed an infinite, number of messages with a small number of signals, a syntax would, in principle, suffice. Without a phonology, however, each smallest unit of an utterance would necessarily be a word, so a talker would have to make do with a very small vocabulary. The obvious function of the phonologic domain is, then, to construct words out of a few meaningless units, and thus to make possible the large vocabularies that human beings like to deploy. But the words of the vocabulary are presumably to be found in the deeper reaches of the phonology, where they are represented by the abstract phonemes that stand beneath the many phonetic variations at the surface, variations associated with phonetic context, word boundaries, rate of articulation, lexical stress, phrasal stress, idiolect, and dialect, to name the most obvious sources. What remains in speaking is, of course, to derive the surface phonetic structures, and then to transmit them by using the organs of articulation to produce and modify sounds. Transmitting those structures as sounds and at high rates becomes the distinctive function of the phonetic mode.

At average rates of speaking, talkers produce and listeners perceive about 8 to 10 segments per second. In the extreme, the rate may go to 25 or 30 per second, at least for short stretches. Plainly, such rates would be impossible if each segment were represented, as in the acoustic alphabets of our early reading machines, by a segment of sound. The organs of the vocal tract cannot make unit gestures that fast, and, even if they could, the rate of delivery of the resulting units of sound would overreach the temporal resolving power of the ear. The trick, then, is to evade the limitations on the rate at which discrete segments of sound can be transmitted and perceived, while yet preserving the discrete phonetic segments those sounds must convey.

The vocal tract solves its part of the problem by breaking the two or three dozen phonetic segments into a smaller number of features, assigning each feature to a gesture that can be made more or less independently, and then turning the articulators loose, as it were, to do what they can. A consequence is that gestures corresponding to features of successive segments are produced at the same time, or else greatly overlapped, according to the constraints and possibilities inherent in the masses to be moved and in the neuromuscular arrangements that move them. This is to say that the character of speech is determined largely by the nature of the mechanisms that do the speaking. But it could hardly be otherwise. For even if Nature had devised articulators that could make successive unit gestures at rapid rates--putting aside that this would presumably have destroyed the utility of the vocal tract for such other purposes as eating and breathing--the resulting drumfire of sound would, as I noted earlier, defeat the ear. At all events, the nature of the articulatory process produces a relation between phonetic segment and sound--the singular code I referred to in the introduction--that must, I think, take first place in any attempt to investigate and understand the perception of speech.

One characteristic of the code that should immediately engage our attention follows from the fact that one or another of the articulators is almost always moving. The consequence is that many, perhaps most, of the potential acoustic cues--that is, aspects of the sound that bear a systematic relation to the phonetic segment--are of a dynamic sort. Witness, for example, the changes in formant frequency, caused by the movement from one articulatory position to another and known to be important cues for various consonants (and, indeed, for vowels) (Liberman, Delattre, Cooper, & Gerstman, 1954; O'Connor, Gerstman, Liberman, Delattre, & Cooper, 1957; Mann & Repp, 1980; Strange, Jenkins, & Edman, 1977). How do these time-varying acoustic cues evoke discrete and unitary phonetic percepts that have no corresponding time-varying quality?

Another characteristic of the code, owing again to the way the articulators produce and modulate the sound, is that the acoustic cues are numerous and diverse. In the contrast between the [b] of <u>rabid</u> and the [p] of <u>rapid</u>, for example, Lisker (1978) has so far identified sixteen cues, representing a variety of acoustic types. The many cues are not ordinarily of equal power--some will override others--but power does not appear to be determined primarily by acoustic prominence. How, then, is such a numerous variety of seemingly arbitrary cues bound into a single phonetic percept?

Finally, the processes of articulation, and more particularly coarticulation, cause the potential cues for a phonetic segment to be widely distributed through the signal and merged, often quite thoroughly, with potential cues for other segments. In a syllable like <u>bag</u>, to take a simple case, it is likely that a single parameter of the acoustic signal--say the second formant--carries information simultaneously about at least two of the constituent segments and, in some places, all three (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952; Liberman, 1974). Indeed, it is this characteristic of speech, this encoding of several phonetic segments into one segment of sound, that is, as we have seen, an essential aspect of the processes by which phonetic segments are produced and perceived at high rates. But the result is an acoustic amalgam, not an alphabet. How does the listener recover from it the string of discrete phonetic segments it encodes?

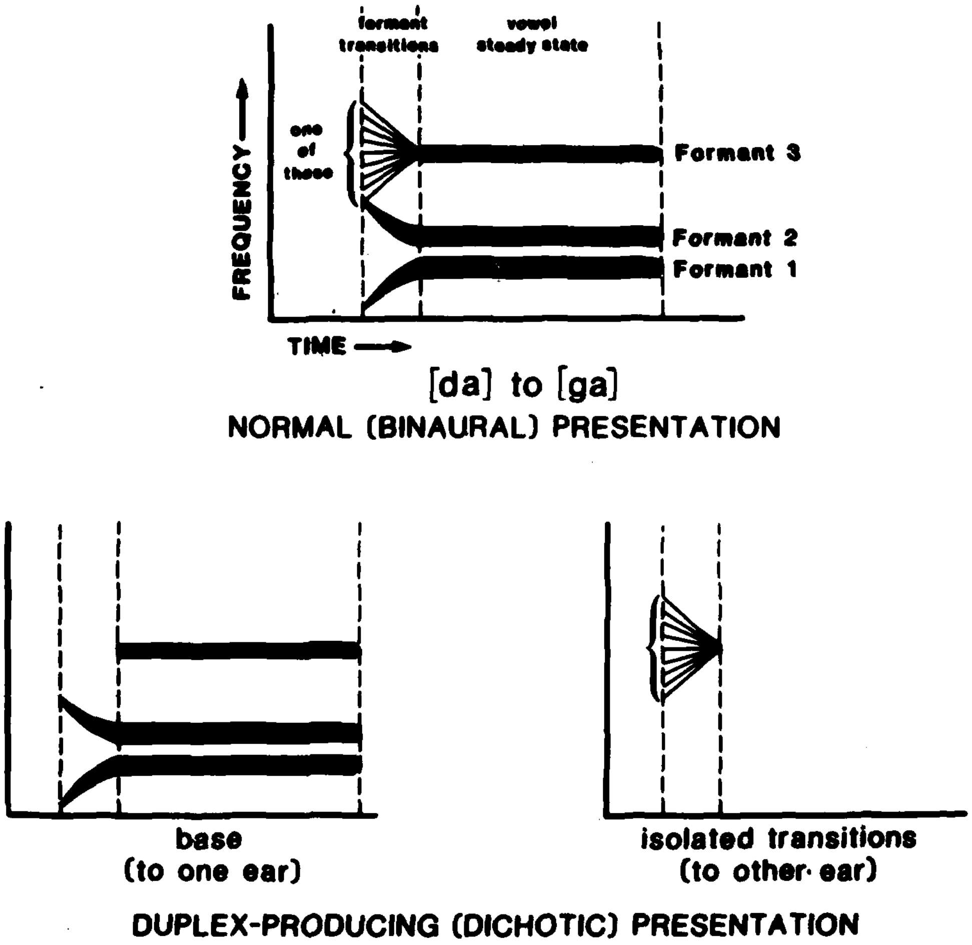
Of course, we might try to evade those questions, and the thorny problems they pose for the auditory mode, by supposing that the articulators produce, for each phonetic segment, at least one cue that represents the segment quite straightforwardly (Stevens & Blumstein, 1981). Because the relation of that cue to the phonetic segment is transparent to ordinary auditory processes, the listener might respond most attentively just to it, dismissing the others as so much chaff, or else learning to accept them as associated with, but wholly *incidental to*, the real business of talker and listener. Such evasion will be hard to maintain, however, if, as we now have reason to think, the typical listener is sensitive to all the phonetic information in speech sounds (Bailey & Summerfield, 1980).[2] Certainly every potential cue so far tested has proved to be an actual cue, no matter how peculiar seeming its relation to the phonetic segment.

We should suppose, then, that there is in speech perception a process by which the manifold of variously merged, continuous, and time-varying cues is made to form in the listener's mind the discrete and ordered phonetic segments that were produced by the speaker. But it seems hardly conceivable that this could be accomplished by processes of a generally auditory sort. Therefore, I assume, as I said in the introduction, that the process is a special one--a *distinctively phonetic process*, specifically adapted to the unique characteristics of the speech code. Since that code is opaque except as one understands the special way it comes about, I find it plausible to suppose, further, that a link between perception and production constrains the process as if by knowledge of what a vocal tract does when it makes linguistically significant gestures (Cooper et al., 1952; Liberman, Delattre, & Cooper, 1952).

A Special Process of the Phonetic Mode: Integration of Cues

Of the many experimental results that bear on the existence and nature of distinctively phonetic processes, none is critical; what tells is the weight of the evidence and the way it converges on certain conclusions. Faced, thus, with many more *results than I could hope to include*, I had to choose between picking a closely related few and, alternatively, offering a token of each type. (For recent and comprehensive reviews, see Repp, 1981; Studdert-Kennedy, 1980). I have chosen the related few, selecting them from recent studies that bear on the three questions raised by the characteristics of the speech code I referred to in the previous section. Aspects of these questions have long been worried about as the problem of "segmentation": how is the acoustic signal "divided" into phonetic segments (Cooper et al., 1952; Fant, 1962; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967)? Recently, Repp (1978) and Oden and Massaro (1978) have looked at the other side of the coin, putting attention on the problem of "integration": how do cues combine to produce the percept? It suits my purposes to adopt their perspective, and so I will.

Integration of a time-varying sound. Frequency sweeps--called formant transitions--of the kind shown in Figure 1 can be sufficient cues for the perceived distinction between the stop consonants [d] and [g] in the syllables [da] and [ga] (Harris, Hoffman, Liberman, & Delattre, 1958). But, as I asked earlier, how are such frequency sweeps integrated (as information about the phonetic dimension of "place") into a unitary percept, [d] or [g], that has about it no hint of a corresponding sweep in pitch? Two interpretations are possible: one, that the integration is accomplished by ordinary auditory processes; the other, that special phonetic processes come into play.



[da] to [ga]
NORMAL (BINAURAL) PRESENTATION

DUPLEX-PRODUCING (DICHOTIC) PRESENTATION

Figure 1. Schematic representation of the stimulus patterns used in the experiment on integration of the time-varying formant transitions (Mann, Madden, Russell, & Liberman, Note 2).

On an auditory interpretation, one might suppose, most simply, that this is an instance of low-level sensory integration, something like the well-known integration of intensity and time into the perception of loudness. That possibility is quickly ruled out, however, by the observation that when the transition cues are removed from the pattern and presented alone, as in the part of the figure at lower right, listeners do perceive a rising or falling "chirp," almost a glissando, that conforms reasonably to the time-varying percept that psychoacoustic considerations might have led us to expect (Mattingly, Liberman, Syrdal, & Halwes, 1971).

But the auditory theory is not so easily disposed of, because it can always fall back on the assumption that the formant transitions collaborate with the rest of the pattern in an interaction of a purely auditory sort, from which the percepts, [d] or [g], emerge. It matters little that there is nothing in what we know about perception of complex sounds to suggest that such interaction should occur, for we know very little about perception of complex sounds. Nor does it necessarily matter how implausible it is to suppose that the articulators could so comport themselves as to produce exactly the right combination of sounds, not just in this instance, but in the myriad others that must occur as the articulators accommodate to variations in, for example, phonetic context, rate, and linguistic stress. Such considerations make an explanation based on auditory interaction endlessly ad hoc, but they do not, in principle, rule it out.

A phonetic interpretation, on the other hand, would have it that the integration of the formant transitions into a unitary percept reflects the operation of a device specialized to perceive the sounds in a linguistically appropriate way. As for what is linguistically appropriate, it is plain that perceiving the transitions as rising or falling chirps is not. Language, after all, has no use for that kind of auditory information; it only requires to know whether the segment was [d] or [g]. Indeed, if the chirps and other curious auditory characteristics of speech sounds were heard as such, they would intrude as an intermediate stage of perception that had, itself, to be interpreted, however automatically. In that case, listening to speech would be like listening to the acoustic alphabets of our early reading machines, or to Morse code. and that would surely be awkward in the extreme.

What is required, if the time-varying transitions are to be perceived (appropriately) as unitary segments, is that the percept reflect neither the proximal sound nor the more distal movements it betokens, but rather the still more distal, and presumably more nearly unitary, neural command structure that occasioned the movements. A less timid writer might call that the talker's phonetic intent.

But whatever the percept exactly corresponds to, I suppose that Nature provided a device that is well adapted to its linguistic function, which is to make available to the listener just those phonetic objects he needs if he is to understand what the speaker said. But Nature could not have anticipated the development of synthetic speech and dichotic stimulation, so it is possible to defeat her design in such a way as to discover something about what the design is. To do this, we use a method that derives from a discovery by Rand (1974). (See also Isenberg & Liberman, 1978; Liberman, 1979). Its special feature is a way of presenting patterns of synthetic speech so that an acoustic cue is perceived as a nonspeech sound and, simultaneously, as support for a phonetic percept. The obvious advantage of the method is that it holds the stimulus input constant while yet producing two percepts, thus providing a control for auditory interaction. Recently, the method has been applied by Mann, Madden, Russell, and Liberman (1981; Note 2) to determine how a time-varying formant transition is integrated into the perception of a stop consonant. The experiment was as follows.

To one ear we presented one or another of the nine formant transitions, as shown at the lower right of Figure 1. By themselves, these isolated transitions sound like time-varying chirps--that is, like reasonably faithful auditory reflections of the time-varying acoustic signal. To the other ear, we presented all the rest of the pattern--the base, so called--that is shown at the lower left of the figure. By itself, the base is always perceived as a stop-vowel syllable; most listeners hear it as [da], some as [ga].

When these two stimuli are presented dichotically, listeners report a duplex percept. On one side of the duplexity, the listeners perceive the syllable [da] or [ga], depending on the identity of the isolated transition. This speech percept is seemingly no different from the one that would have been produced had the base and the isolated transition been electronically mixed and presented in the normal manner. On the other side, and at the same time, the listeners perceive a nonspeech chirp, not perceptibly different from what they experience when the transition is presented by itself. Thus, given exactly the same acoustic context, and the same brain, the transition is simultaneously perceived in two phenomenally different ways: as critical support for a stop consonant, in which case it is integrated into a unitary percept, and as a nonspeech chirp, in which case it is not.

To go beyond the phenomenology just described, we determined how the transitions would be discriminated, depending on which side of the duplex percept the listener was attending to. For that purpose, we sampled the continuum of formant transitions by pairs, choosing, as members of each to-be-discriminated pair, stimuli that were three steps apart on the continuum of formant transitions shown in Figure 1. These we presented in an AXB format (A and B being the two stimuli to be discriminated and X being the one or the other) to subjects who were instructed to decide on the basis of <u>any</u> perceptible difference whether X was more like A or like B. When the subject's attention was directed to the speech side of the duplex percept, we obtained results represented in Figure 2 by the solid line; with attention directed to the nonspeech side, we obtained the results shown by the dashed line. The difference is obvious. When the transitions support stop consonants--that is, when they are perceived in the phonetic mode--the discrimination function has a rather high peak, the location of which corresponds closely to the phonetic boundary. This is the familiar tendency toward categorical perception that characterizes segments such as these, a tendency that is, itself, rather highly adaptive, since it is only the categorical information--the segment is categorically [d] or [g]--that is most relevant linguistically. When the same transitions are perceived, on the nonspeech side of the percept, as chirps, the discrimination function, shown as the dashed line and open circles, is different; in fact, it is nearly continuous.[3] Thus, the discrimination functions confirm the more blatantly phenomenological results described earlier. Both indicate that integration of the formant

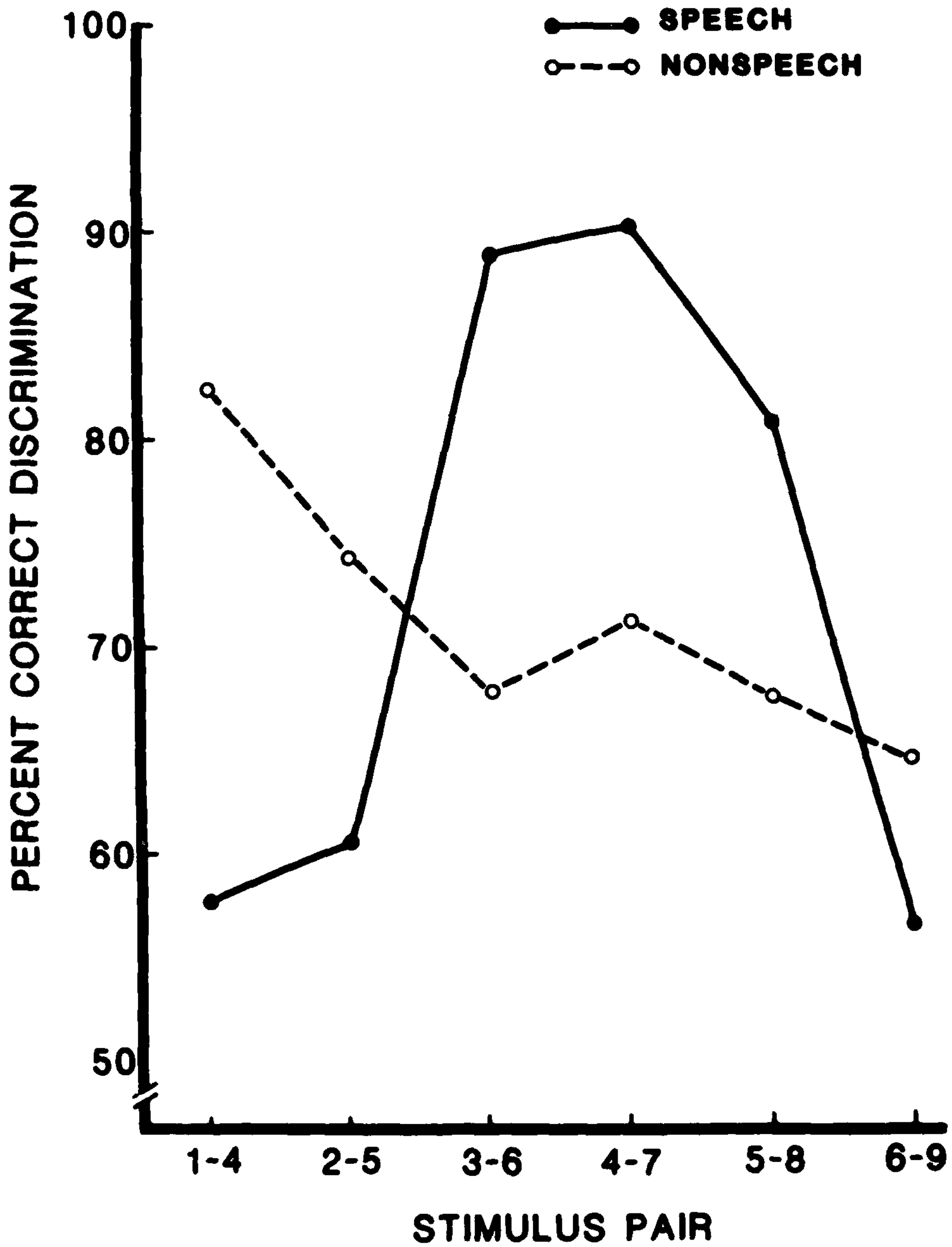


Figure 2. Discriminability of formant transitions when, on the speech side of the duplex percept, they supported perception of stop consonants, and when, on the nonspeech side, they were perceived as chirps (from Mann, Madden, Russell, & Liberman, Note 2).

transition into a phonetic percept is owing to a special process that makes available to perception a unitary phonetic object well suited to its role in language.

The same phonetic process that integrates the transitions has other characteristics, of course, including one that has attracted attention for a long time: it adjusts perception to variations in the acoustic signal when those are caused by coarticulatory accommodation to changes in phonetic context; thus, it seems to rest on a link between perception and production (Liberman et al., 1952; Mann, 1980; Mann & Repp, 1981). A second part of the experiment just described was designed to examine that perceptual adjustment to phonetic context, and to exploit the duplex percept to identify the domain, auditory or phonetic, in which it occurs. To that end, we took advantage of an earlier experiment by Mann (1980) in which she had found that placing the syllables [al] or [ar] in front of the [da]-[ga] patterns caused the position of the [da]-[ga] boundary (on the continuum of formant transitions) to shift--toward the [g] end for [ar] and the [d] end for [al]. Since the shift was consistent with the change in [da]-[ga] articulation that can be shown to occur when the syllable [al] or [ar] is spoken immediately before, Mann inferred that this was, indeed, a case in which the perceptual system had automatically reflected coarticulation and its acoustic consequences.

Our further contribution to Mann's result was simply to repeat her experiment, but with the "duplex" procedure (and with measures of discrimination substituted for the phonetic identifications she had used). The outcome was quite straightforward. On the speech side of the duplex percept we (in effect) replicated the earlier result, as shown by the results displayed in Figure 3. Taking the discrimination data obtained with the isolated [da]-[ga] syllables (solid line connecting solid circles) as baseline, we see that placing the syllable [ar] in front caused the discrimination peak (and presumably the phonetic boundary) to move to the right, toward the [g] end of the continuum of transitions. When [al] preceded, the peak (and the boundary) apparently shifted in the opposite direction--that is, to the left, toward [d]; for some subjects, indeed, it shifted so far as to move off the stimulus continuum, so there is, for them, no effective boundary, which explains why the peak is so low. For present purposes, however, the point is simply that there are large effects of prior phonetic context on discrimination of the transitions when those are perceived on the speech side of the duplex percept. On the other hand, as we see in Figure 4, the nonspeech side of the percept is unaffected by phonetic context: discrimination of the formant transitions is the same whether the base was preceded by [al], by [ar], or by nothing.

Putting the two experiments together, we conclude that, given a single acoustic context, exactly the same formant transitions are perceived in two different modes. In the one mode, they evoke nonspeech chirps that have a time-varying quality corresponding, approximately, to the time-varying stimulus; changes in the transitions are perceived continuously; and perception is unaffected by phonetic context. This is, of course, the auditory mode. In the other mode, the same transitions provide critical support for the perception of stop consonants that lack the time-varying quality of the nonspeech chirps; changes in the transitions are perceived more or less categorically; and perception is markedly affected by phonetic context. This is the phonetic mode.

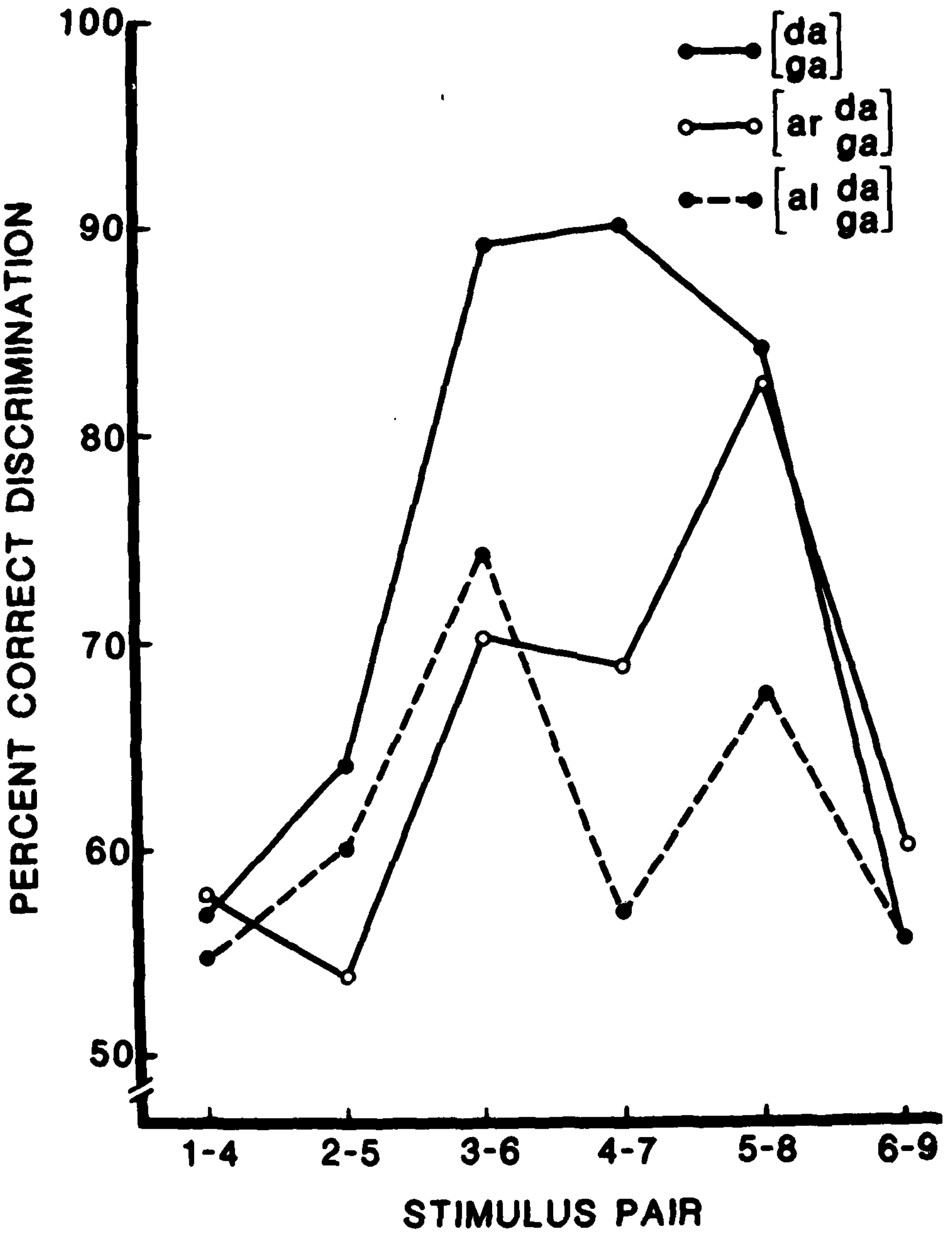


Figure 3. Discriminability of the formant transitions on the speech side of the duplex percept when the target syllables [da] and [ga] were in *isolation and* when they were presented by the syllables [ar] and [al] (from Mann, Madden, Russell, & Liberman, Note 2).

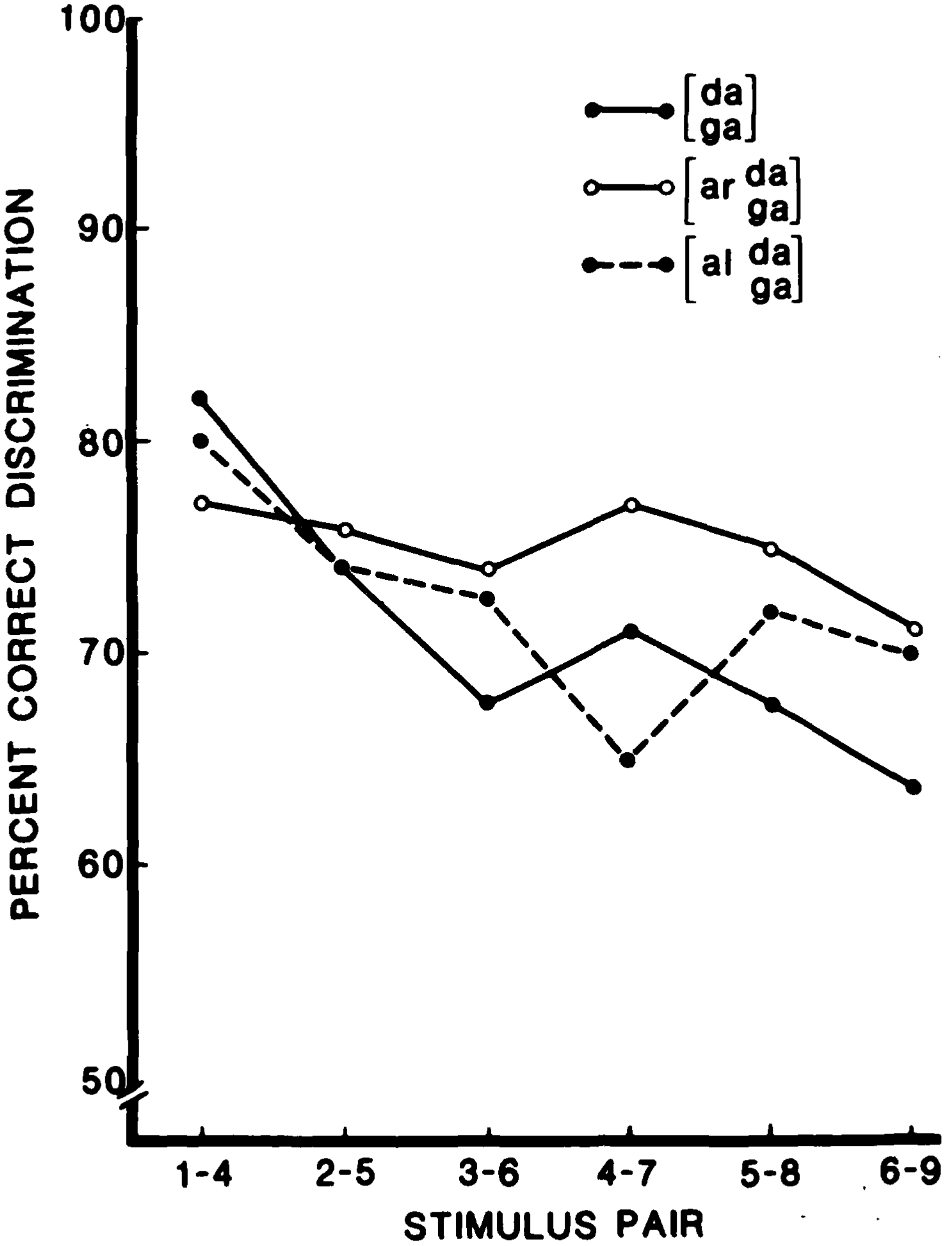


Figure 4. Discriminability of the formant transitions on the nonspeech side of the duplex percept under conditions identical to those represented in Figure 3 (from Mann, Madden, Russell, & Liberman, Note 2).

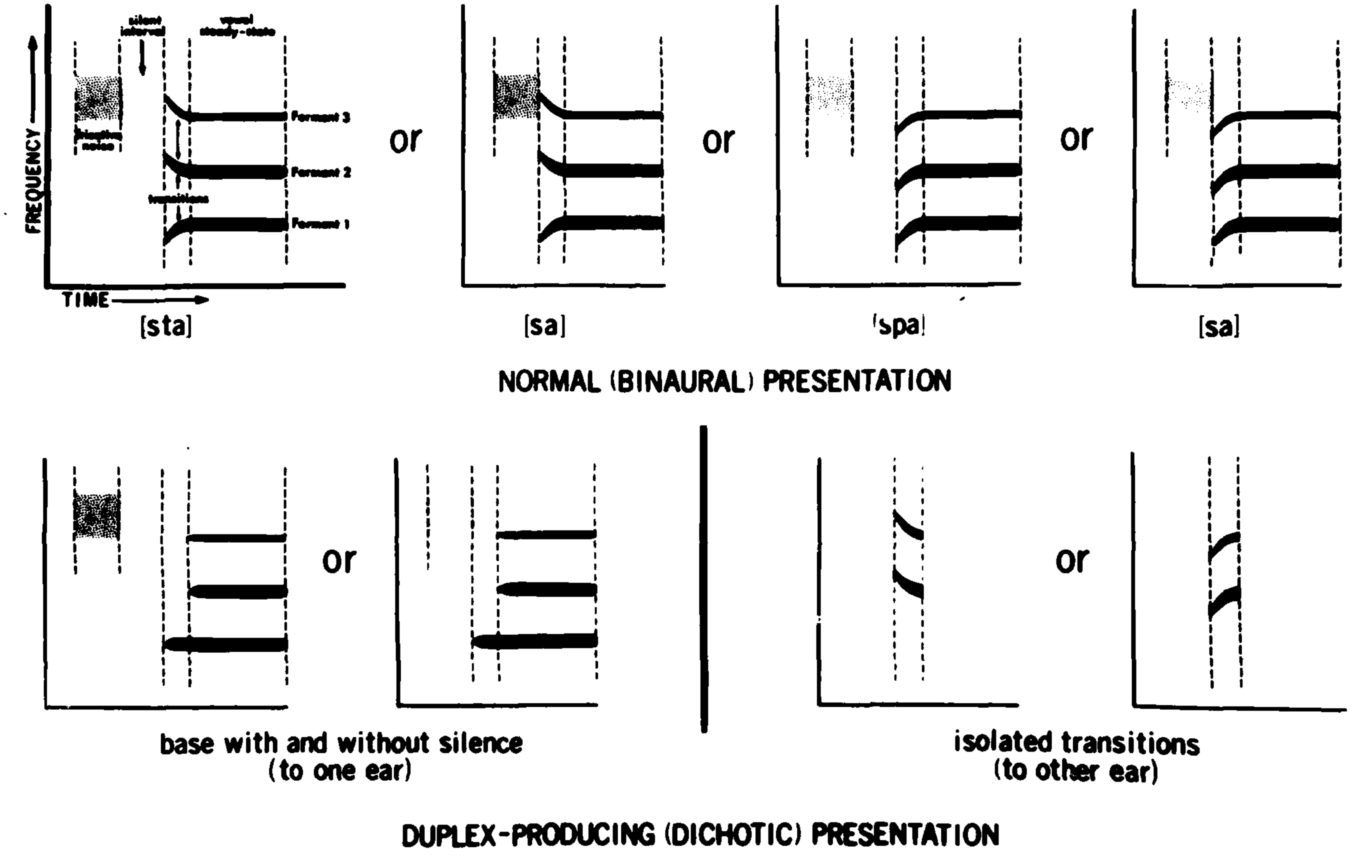


Figure 5. Schematic representations of the stimulus patterns used to determine whether the importance of silence as a cue is owing to auditory or phonetic factors. (From "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," by A. M. Liberman, D. Isenberg, and B. Rakerd, *Perception & Psychophysics*, in press. Copyright by the Psychonomic Society, Inc. Reprinted by permission.)

Integration of sound and silence. Perception of a phonetic segment typically depends, as I indicated earlier, on the integration of several--many may be a more appropriate word--acoustic cues. Even in the case of [da] and [ga] just described, there was one other cue, silence preceding the transitions, though I did not remark it. To show the effect of such silence--an effect long known to researchers in speech (Bastian, Delattre, & Liberman, 1959)--we must put the stop consonant and its transition cues into some other position, as in the examples [spa] and [sta] shown at the top of Figure 5. As we see there, an important cue for perception of stop consonants--in this case, [p] and [t]--is a short period of silence between the noise of the fricative and the formant transitions that introduce the vocalic part of the syllable (Dorman, Raphael, & Liberman, 1979).

But why is silence necessary, and in which domain, auditory or phonetic, is it integrated with the transition cues to produce stop consonants? On an auditory account, we might suppose that there is forward masking of the transition cues by the fricative noise, in which case the role of the intervening silence is to provide time for the transitions to evade masking. Failing that, we could, as always, invoke some previously unnoticed interaction between frequency sweeps (transitions) and silence that is presumed to be characteristic of the way the auditory system works.

A phonetic interpretation, on the other hand, takes account of the fact that presence or absence of silence supplies important phonetic information--to wit, that the talker closed his vocal tract, as he must to produce the [p] and [t] in [spa] and [sta], or that he did not, as he does not when he says [sa]. Presumably, the processes of the phonetic mode are sensitive to the phonetic significance of the information that silence imparts.

To decide between these interpretations, the phenomenon of duplex perception was again exploited (Liberman, Isenberg, & Rakerd, in press). As shown in Figure 5, base stimuli that sometimes did, and sometimes did not, have silence were presented dichotically with transition cues appropriate for [p] or for [t]. Two such dichotically yoked patterns were presented on each trial; subjects were asked to identify the speech percepts and to discriminate the nonspeech chirps. The result was that the subjects fused the transitions with the base and accurately perceived [sa], [spa], or [sta], depending on the presence or absence of silence in the base (to one ear) and the nature of the formant transitions (to the other). But the subjects also perceived the transitions as nonspeech chirps, and accurately discriminated them as same or different regardless of whether or not there was silence in the base. Thus, duplex perception did occur, and silence affected the identification of the speech, but not the discrimination of the nonspeech.

In a further experiment, the investigators provided a more severe test by asking subjects to discriminate their percepts on both sides of the duplexity. For that purpose, two dichotically yoked pairs of stimuli were presented on each trial, so arranged as to exhaust all combinations of silence-no silence in the base and [p]-[t] cues in the isolated transitions. Subjects were asked, for each pair of percepts, to rate their confidence that a difference of any kind had been detected. The results are shown in Figure 6. There are but two critical comparisons. The first is in the leftmost third of the figure, in the condition in which there was no silence in either of the two

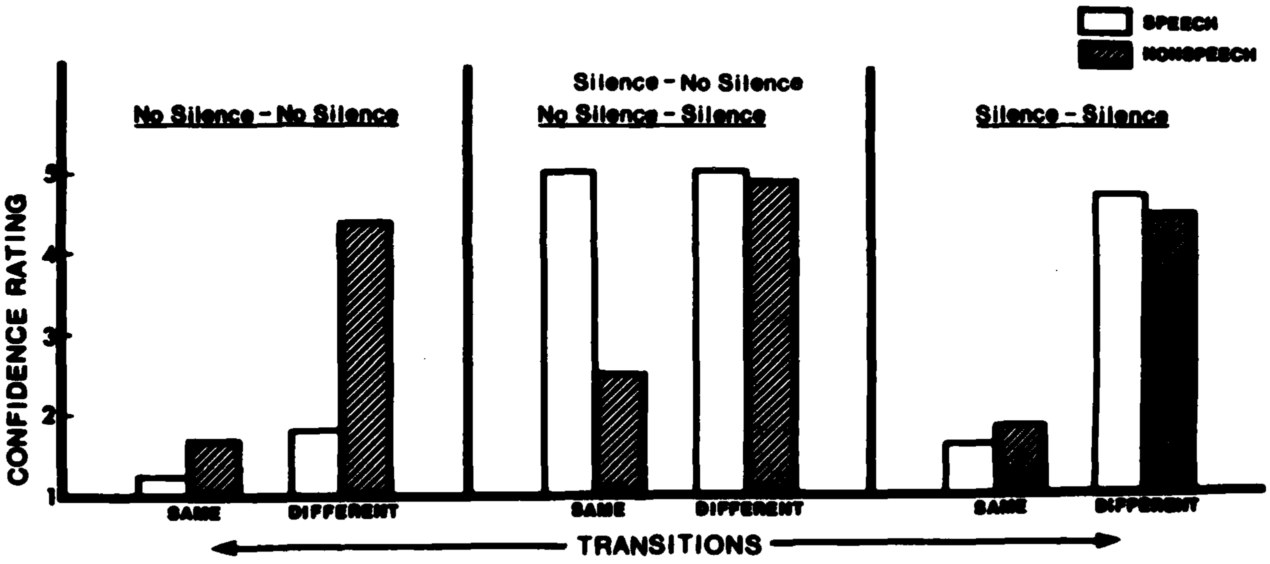


Figure 6. Mean ratings of confidence that the two percepts (speech or nonspeech) were different. (From "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," by A. M. Liberman, D. Isenberg, and B. Rakerd, Perception & Psychophysics, in press. Copyright by the Psychonomic Society, Inc. Reprinted by permission.)

base stimuli presented to the one ear (labelled "No Silence - No Silence") and the two transition cues to the other ear were different (labelled simply "Different"). On the speech side of the duplexity (open bar), we see that the difference between the transitions was not clearly detected, presumably because, in the absence of silence in either base stimulus, subjects perceived [sa] in both cases. But, on the nonspeech side (shaded bar), the same difference was detected; here, the absence of silence in the base made no difference. The other critical comparison is seen in the bars immediately to the right, in the middle third of the slide, representing the condition that had, in the one ear, silence in one base stimulus but not the other, and, in the other ear, two transition cues that were the same. On the speech side of the duplex percept, we see that the patterns were perceived as very different, even though the transition cues were the same; presumably, this was because one percept, being influenced by the presence of silence, included a stop consonant, while the other, being influenced by the absence of silence, did not. The result on the nonspeech side stands in contrast. There, the percepts were judged to be not very different, accurately reflecting the fact that they were, in fact, not different.

Thus, in both critical comparisons, silence affected discrimination of the transitions only on the speech side of the duplex percept. Apparently, its importance depends on distinctively phonetic processes, and its integration with the transition occurs in the phonetic mode.

The integration of silence and transitions, as in the patterns just described, reinforces the suggestion, made earlier in regard to the integration of the transitions alone, that the perceived object is not to be found in the movements of the speech organs at the periphery, but rather at some still more distal remove, as suggested by Repp, Liberman, Eccardt, and Pesetsky (1978). To see the point more clearly, we should first take note of a finding that adds another cue for the [p] in [spa]: the shaping of the fricative noise that is caused by the way the vocal tract closes for [p] (Summerfield, Bailey, Seton, & Dorman, 1981). Now we have three acoustic cues that correspond neatly to three corresponding aspects of the articulation. There is, first, the shape of the fricative noise, which signals the closing of the tract; then the silence, which signals the closure itself; and finally the formant transitions, which signal the subsequent opening into the vowel. If these three acoustic cues are integrated into a percept that does not display at least three constituent elements, then the perceived object must be upstream from the peripheral articulation. A likely candidate, as suggested earlier, is the unitary command structure from which the various movements at the periphery unfolded.

Integration of periodic sound and noise. When a talker closes his vocal tract to produce a stop consonant and then opens it into a following vowel, the resulting silence and formant transitions are, as we have seen, integrated into a stop consonant. It is surely provocative that similar formant transitions are produced, but without the silence, when a talker almost closes his vocal tract so as to make the noise of a fricative (e.g., [s]), and then opens into the vowel, for in such cases the formant transitions do not support stops; they are, instead, integrated with the noise into the perception of a fricative (Harris, 1958; Mann & Repp, 1980; Whalen, 1981). Such integration is shown in Figure 7, where I have reproduced the results of a recent

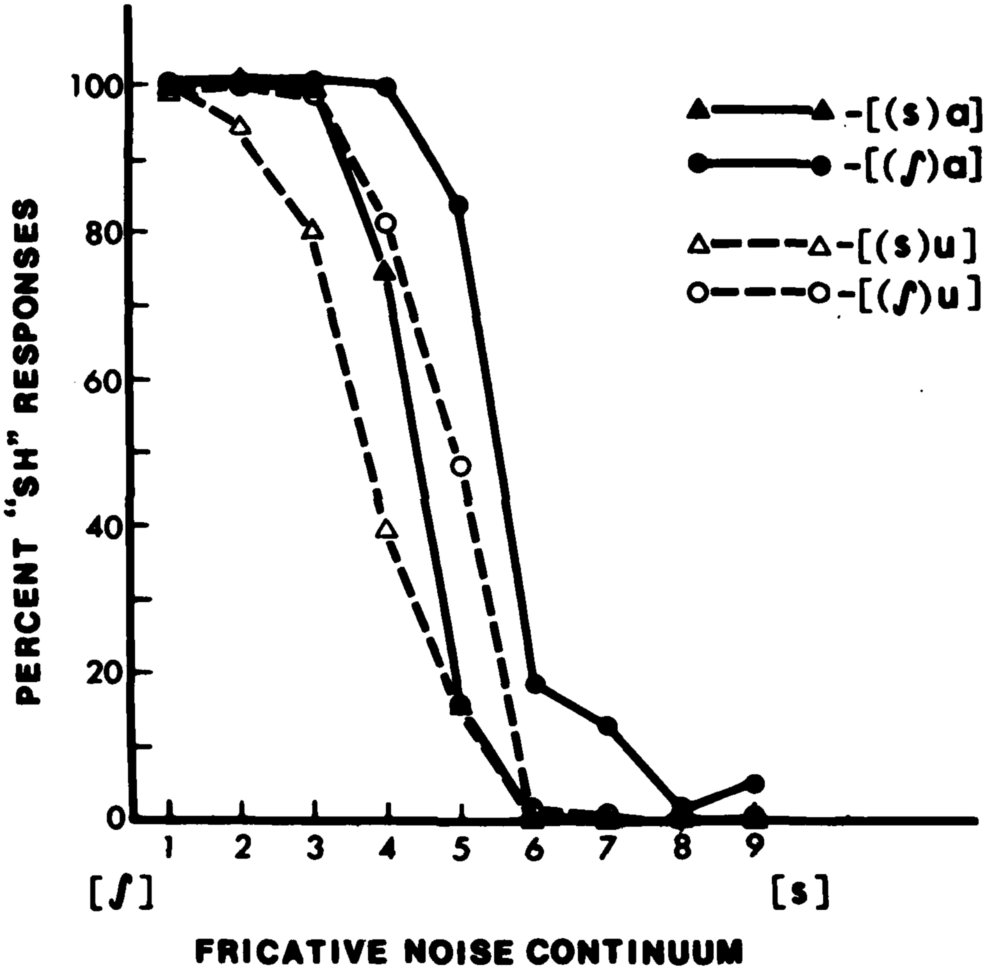


Figure 7. Identification functions for a [ʃ] -[s] noise continuum when connected to [ʃ]-appropriate or [s]-appropriate transitions and the vowels [a] or [u]. (From "Two strategies in fricative discrimination," by B. H. Repp, Perception & Psychophysics, in press. Copyright by the Psychonomic Society, Inc. Reprinted by permission.)

experiment by Repp (in press). What we see in the figure are the judgments [ʃ] or [s] made to stimuli that were constructed as follows. The experimental variable, ranged on the abscissa, was the position on the frequency scale of a patch of band-limited noise as it moved between a place appropriate for [ʃ] and one appropriate for [s]. The parameters were the nature of the (following) formant transitions--appropriate, in the one case, for [s] and, in the other, for [ʃ]--and the two vowels [a] and [u]. We see that the transitions (and also the vowels) affected the perception of the fricative.

Though not shown in this particular experiment, I would note, parenthetically, that patterns like these, but with 50 msec of silence inserted between the fricative noise and the vocalic section, will be perceived, not as fricative-vowel syllables, but as fricative-stop-vowel syllables (Mann & Repp, 1980). That is, inserting 50 msec of silence will cause the formant transitions to be integrated, not into fricatives, but into stops. It is difficult to account for that as an auditory effect, but easy to see how it might reflect a special sensitivity to information about a difference in articulation that changes the phonetic "affiliation" of the acoustic transitions.

In a further, and more severe, test of the integration of transitions and fricative noise that we saw in Figure 7, Repp measured the effect of the formant transitions on the way listeners discriminated variations in the frequency position of the noise patch, using for this purpose the highly sensitive method of "fixed standard." He found two distinctly different types of discrimination functions. One clearly showed an effect of the formant transitions and reflected nearly categorical perception; the other just as clearly showed no effect of the formant transitions and represented perception that was nearly continuous. Which type Repp obtained in each particular case depended, apparently, on the listener's ability to isolate or "stream" the noise--that is, to create an effect similar, perhaps, to the one obtained by Cole and Scott (1973) when they found with fricative-vowel syllables that, as a result of repeated presentation, the noise and vocalic sections would form separate "streams" that had little apparent relation to each other. At all events, we have here another instance, though occurring in a different phonetic class and obtained by very different methods, of a single acoustic pattern that is perceived in two distinctly different ways. One reflects the integration of cues in the phonetic mode, the other the "nonintegration" of the same acoustic elements in the auditory mode.

There is still another method that exploits the possibility of perceiving exactly the same stimulus pattern in two ways, and thus enables us to test yet again whether the integration of formant transitions and noise occurs in the phonetic or auditory modes. But, now, the two ways of perceiving are not speech versus nonspeech, as in the experiments described thus far, but rather two kinds of speech--namely, fricatives and stops. The relevant experiment is a recent one by Carden, Levitt, Jusczyk, and Walley (1981). Starting with synthetic patterns that produced stop-vowel syllables, they varied the second-formant transitions and found the boundary between [b] and [d]. Then they placed in front of these patterns a fixed patch of band-limited noise, neutralized as between the fricatives [f] and [θ]. In these patterns, the formant transitions cue the difference between the fricatives, but, because the place of vocal-tract constriction is different for the two fricatives, on

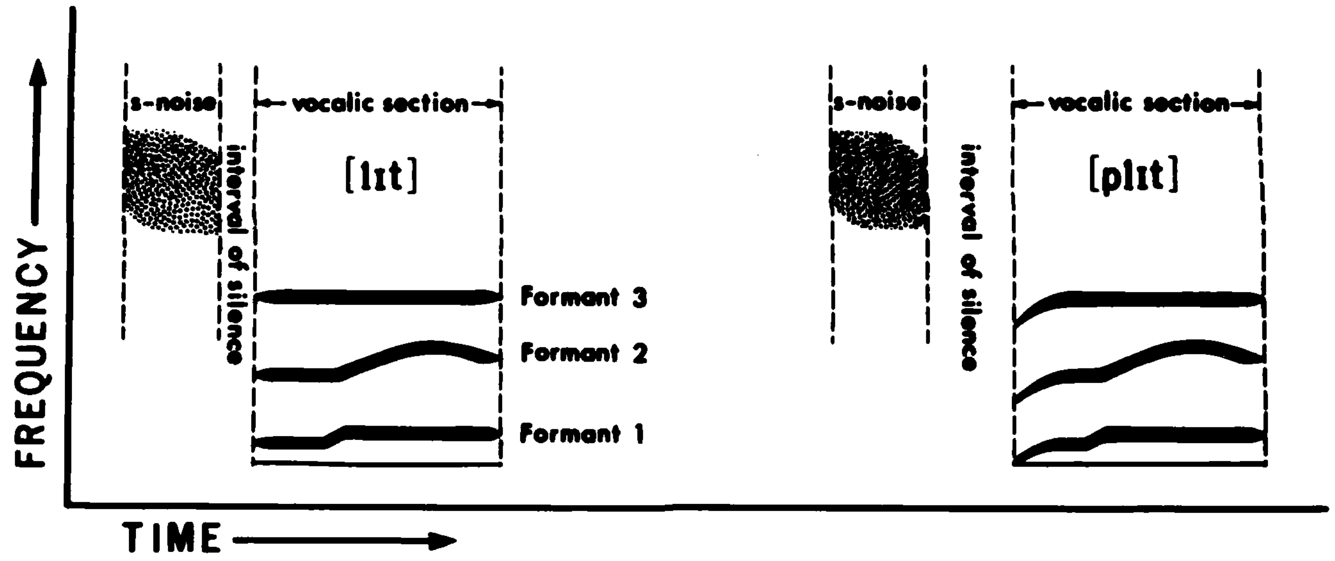


Figure 8. Schematic representation of the patterns used to evaluate the equivalence in stop-consonant perception of silence and formart transitions, showing both settings of the transitions and two representative settings of the silence cue. (From "Perceptual equivalence of two acoustic cues for stop-consonant manner," by H. L. Fitch, T. Halwes, D. M. Erickson, and A. M. Liberman, Perception & Psychophysics, 1980, 27, 343-350. Copyright 1980 by the Psychonomic Society, Inc. Reprinted by permission.)

the one hand, and the two stops, on the other, the perceptual boundary on the continuum of formant transitions is now displaced. That is, exactly the same formant transitions distinguish the fricatives differently from the way they distinguish the stops. The effect seems most plausibly to be phonetic, reflecting the listener"s "knowledge," as it were, of the difference in articulatory place of production between the stops, [b] and [d], on the one hand, and the fricatives, [f] and [ ], on the other. But, just to make sure, Carden and his collaborators presented the patterns with the noise patch to one group of subjects and boldly asked them to perceive stops; then, in precisely reverse fashion, they presented the patterns without the noise patch to a second group with instructions to perceive fricatives. The listeners' judgments reflected boundaries on the continuum of transitions that were appropriate to the class of phonetic segments ([b] vs. [d] or [f] vs. [θ]) they were asked to hear. Thus, exactly the same acoustic patterns yielded different boundaries on the continuum of transitions, depending on whether the listeners were perceiving the patterns as stops or as fricatives. Discrimination functions were also obtained, and these confirmed the boundary shift. We see, then, that transition cues like those that integrate with silence to produce a stop consonant will integrate with noise to produce a fricative. In both cases, the integration is in the phonetic mode.

The equivalence of sound and silence when integrated. Implicit in the discussion so far is the assumption that when acoustic cues integrate to form a phonetic percept, they are, for that purpose, perceptually equivalent; otherwise, it would make no sense to speak of the percept as unitary. It is not implied that the cues are necessarily of equal importance or power, only that their separate contributions are not sensed as separate. But even that implication is of interest from a theoretical point of view, because the cues are often very different acoustically, having in common only that they are the common products of the same linguistically significant gesture. Hence their equivalence is to be attributed, most reasonably, to the link between perception and production that presumably characterizes phonetic processes.

But the implied equivalence of diverse cues is so far just that--implied. To test the equivalence more directly was the purpose of several experiments. One of these, by Fitch, Halwes, Erickson, and Liberman (1980), was designed to examine the equivalence of silence and formant transitions in perception of the stop consonant in *split* as opposed to its absence in *slit*. Synthetic patterns like those shown in Figure 8 were used. The variable was the duration of silence between the fricative noise and the vocalic portion of the syllable; the parameter of the experiment was the nature of the formant transitions at the start of the vocalic sections, set so as to bias that section toward [lit], in the one case, and toward [plit] in the other. When stimuli that had been constructed in this way were presented for identification as *slit* or *split*, the results shown in Figure 9 were obtained. One sees there a trading relation not different in principle from those found by other investigators with other cues. (For a review, see, again, Repp, 1981). The displacement of the two response functions indicates that, for the purpose of producing the [p] in *split*, about twenty msec of silence is equal to appropriate formant transitions.[4] Thus, silence is equivalent to sound, but only, I should think, when both are produced as parts of the same phonetic act.

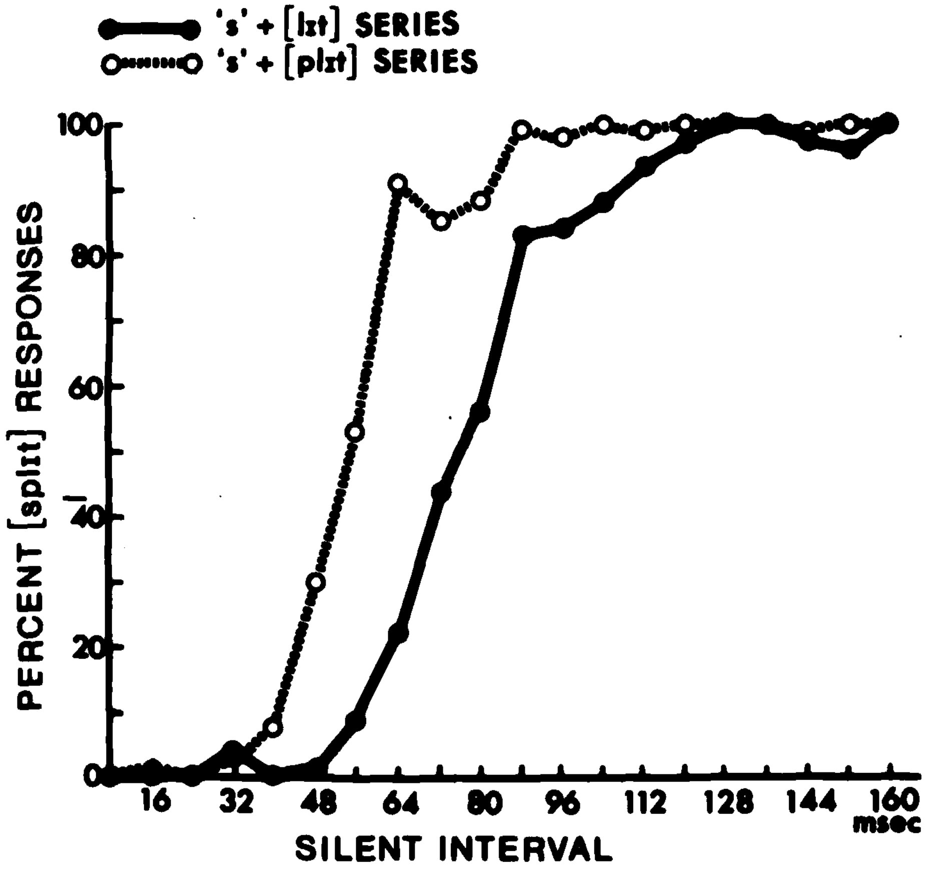


Figure 9. Effect of silent interval on perception of /slit/ vs. /split/ for the two settings of the transition cue. (From "Perceptual equivalence of two acoustic cues for stop-consonant manner," by H. L. Fitch, T. Halwes, D. M. Erickson, and A. M. Liberman, Perception & Psychophysics, 1980, 27, 343-350. Copyright 1980 by the Psychonomic Society, Inc. Reprinted by permission.)

Of course, it might be argued that the splits produced by the two different combinations of silence and sound were not really equivalent, but the forced-choice identification procedure, permitting only the responses slit or split, gave the subjects no opportunity to say so. Against that possibility, we carried out another experiment, designed to determine how well the subjects could discriminate selected combinations of the stimuli on any basis whatsoever. The rationale for selection of stimuli was as follows. If the two cues, silence and sound, are truly equivalent in phonetic perception, their perceptual effects should be algebraically additive, as it were. Thus, given two synthetic syllables to be discriminated, and given a base-line level of discriminability determined for pairs of stimuli that differ in only one of the cues, it should be possible to add the second cue so as to increase or decrease discriminability, according as the phonetic "polarity" of the two cues causes their effects to work together or at cross purposes. The cues should "summate," or "cooperate," when they are biased in the same phonetic direction--as when one of the syllables to be discriminated combines a silence cue that is longer by the amount of the "trade" with transition cues of the [plit] type, and the other syllable combines a silence cue that is shorter by the amount of the "trade" with transition cues of the [lit] type. They should "cancel" each other or "conflict," when the opposite pairing is made--that is, when the longer silence cue is combined with transition cues of the [lit] type, and the shorter silence cue with transition cues of the [plit] type. Pairs of stimuli meeting those specifications, and sampling the continuum of silence durations, were presented for forced-choice discrimination. As shown in Figure 10, discrimination of patterns differing by both cues was, in fact, either better or worse than patterns that differed by only one, depending on whether the cues were calculated to "cooperate" or to "conflict." Apparently, the effects of the two cues did converge on a single perceptual object. By this test, then, the cues may be said to be equivalent and the percept may be said to be truly unitary.

That the equivalence of silence and sound in the above example is owing to phonetic processes is supported in an experiment by Best, Morrongiello, and Robson (1981). Indeed, it is supported there more strongly than in the experiment just described, because Best and her collaborators found that the equivalence was manifest only when the stimulus patterns were perceived as speech. As a first step, they performed an experiment very similar to the one by Fitch et al., except that the stimuli were say-stay instead of slit-split, and the transition-cue parameter was simply the frequency at which the first formant started. With these stimuli, they obtained the identification functions shown in Figure 11. We see there almost exactly the same kind of trading relation between silence and formant transition that had been found in the earlier experiment. In the manner of Fitch et al., they also tested discrimination, finding, just as Fitch et al. had, that the two cues could be made to cooperate or to conflict depending on their phonetic polarities. But now they performed an experiment that proved to be particularly revealing. Borrowing a procedure that had been used successfully for a similar purpose (Lane & Schneider, Note 3; Bailey, Summerfield, & Dorman, 1977; Dorman, 1979), and more recently made the object of further attention (Remez, Rubin, Pisoni, & Carrell, 1981), they replaced the formants of the vocalic portion of the syllable with sine waves, taking care that the sine waves followed exactly the course of the formants they replaced. The sounds that result are perceived by most people, at least initially, as nonspeech patterns of noises and tones.

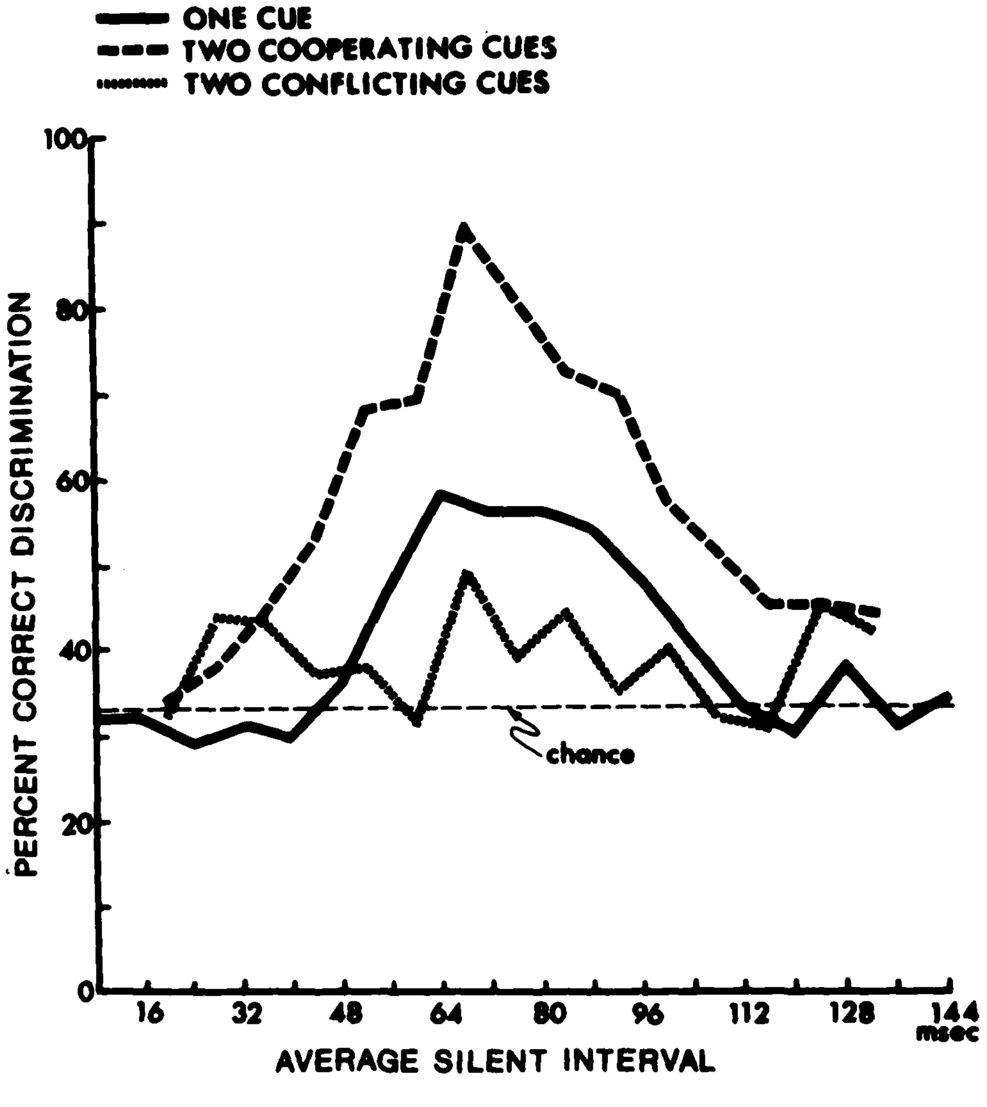


Figure 10. Percent correct discrimination for pairs of stimuli that differ by one cue or by two cues of the same (cooperating cues) or opposite (conflicting cues) phonetic polarities. (From "Perceptual equivalence of two acoustic cues for stop-consonant manner," by H. L. Fitch, T. Halwes, D. M. Erickson, and A. M. Liberman, Perception & Psychophysics, 1980, 27, 343-350. Copyright 1980 by the Psychonomic Society, Inc. Reprinted by permission.)

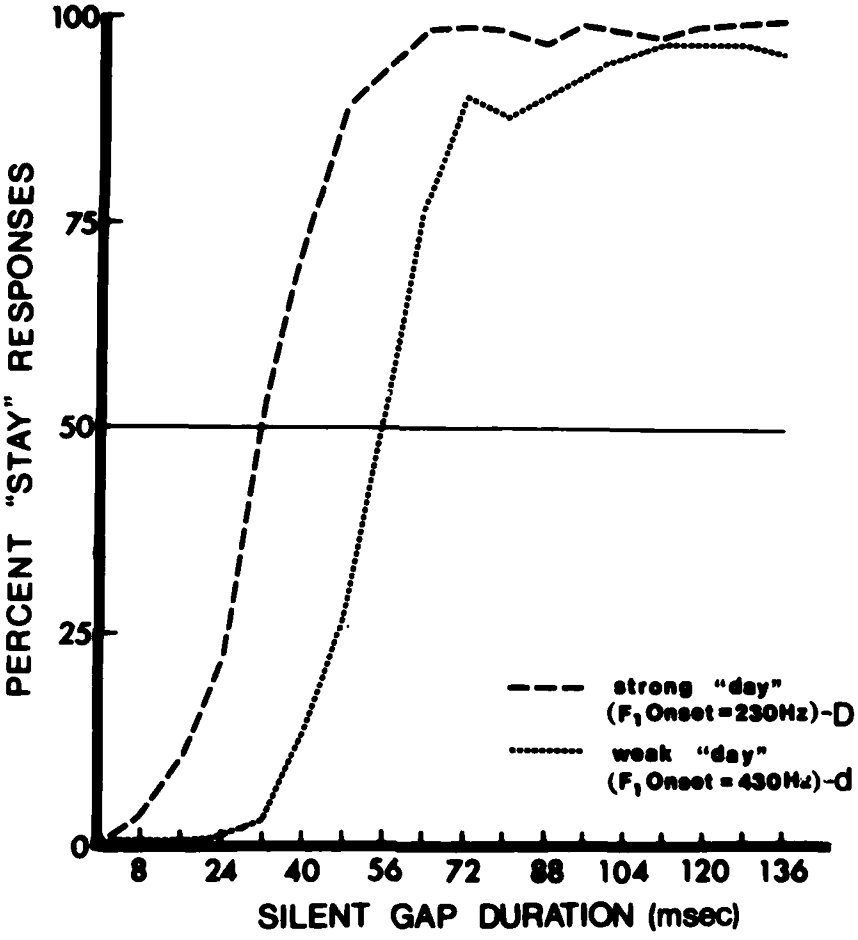


Figure 11. Effect of silent interval on perception of /say/ vs. /stay/ for the two settings of the transition cue. (From "Perceptual equivalence of acoustic cues in speech and nonspeech perception," by C. T. Best, B. Morrongiello, & R. Robson, Perception & Psychophysics, 1981, 29, 191-211. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

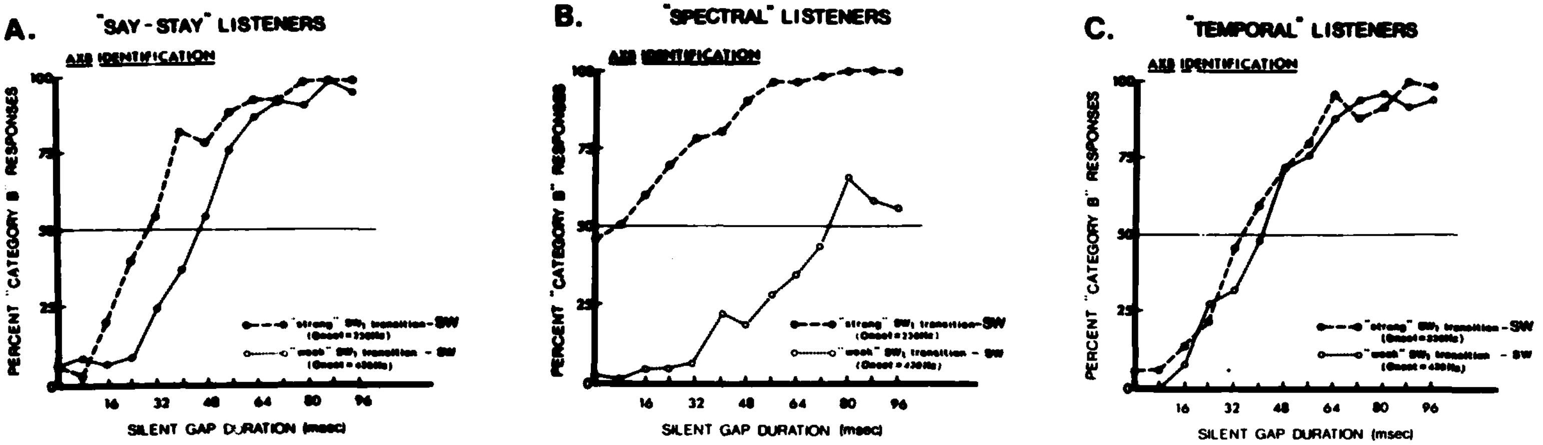


Figure 12. Effect of silent interval on "identification" of sine-wave analogues of say-stay stimuli. Graph A is for those subjects who perceived these stimuli as speech ("say-stay" listeners). Graphs B and C are for those who perceived them as nonspeech, divided, according to their reports of what the sounds were like, into those who were apparently attending to the transition cue (Graph B, "spectral" listeners) or, alternatively, the silence cue (Graph C, "temporal" listeners). (From "Perceptual equivalence of acoustic cues in speech and nonspeech perception," by C. T. Best, B. Morrongiello, and R. Robson, Perception & Psychophysics, 1981, 29, 191-211. Copyright 1981 by the Psychonomic Society, Inc. Reprinted by permission.)

But some spontaneously perceive them as speech, and others perceive them so after it has been suggested to them that they might. It is possible, thus, to obtain identification and discrimination functions for the same stimuli when, in the one case, they are perceived as speech and when, in the other, they are not. (When perceived as nonspeech the patterns are, of course, not readily identifiable, but identification functions can be obtained by presenting, on each trial, the target stimulus--that is, the stimulus to be identified--together with the two stimuli at the extremes of the continuum, and then asking the subject to say whether the target stimulus is more like one or the other of the extremes. To insure comparability, the same procedure is used when the subjects are perceiving the stimuli as speech.) The results are shown in Figure 12. We see, in Figure 12a, that when the subjects were perceiving the patterns as speech ("say-stay" listeners), their identification functions exhibited the now familiar trading relation. But when the same stimuli were perceived as nonspeech, then, as shown in Figures 12b and 12c, two quite different patterns emerged, depending on whether, as inferred from the subjects' descriptions of the sound, they were attending to the transition cue ("spectral" listeners) or the silence cue ("temporal" listeners). It is, of course, precisely because the subjects could not integrate the cues in the nonspeech percept that they chose, as it were, between the one cue and the other. In any case, both of the identification functions in the nonspeech case are different from the one that characterizes the response to exactly the same stimuli when they were perceived as speech. (Discrimination functions obtained with the same stimuli were also different depending on whether or not the stimuli were perceived as speech, nicely confirming the result obtained with the identification measure.) Thus, with yet another method for obtaining speech and nonspeech percepts from the same stimulus, we again find evidence supporting the existence of a phonetic mode, and we see that the equivalence of integrated cues is to be attributed to the distinctively phonetic processes it incorporates.

The equivalence of sound and sight when integrated. Perhaps the most unusual evidence relevant to the issue I have been discussing comes from a startling discovery by McGurk and MacDonald (1976) about the influence on speech perception of optical information about the talker's articulation. (See also MacDonald & McGurk, 1978; Summerfield, 1979). When subjects view a film of a talker saying one syllable, while a recorded voice says another, then, under certain conditions, they experience a unitary percept that overrides the conflicting optical and acoustic cues. Thus, for example, when the talker articulated [ga] or [da] and the voice said [ba], most subjects perceived [da]. In that case, the effect of the optical stimulus was, at the very least, to determine place of production. When, in a subsequent experiment by McGurk and Buchanan (Note 4), the talker was seen to produce the syllables [ba], [va], [ða], [da], [ʒa], [ga], [ha], while the recorded voice said [ba] over and over again, most subjects perceived [ba], [va], [ða], [da], [da], and then, for visual [ha], a variety of percepts other than [ba]. Here, both place of articulation and manner of articulation were determined by the optical input. (The difficulty of seeing farther back in the vocal tract than [da] accounts, presumably, for the fact that visual [ʒa], [ga], and [ha] were perceived as having generally more forward places of production.)

Having witnessed a demonstration of the McGurk-MacDonald effect, I take the liberty of offering testimony of my own. I found the effect compelling, but, more to the point, I would agree that McGurk and Buchanan (Note 4) have captured my experience when they say, "...the majority of listeners have no awareness of bimodal conflict ...," and then describe the percept as "unified." Surely, my percept was unified in the important sense that I could not have decided by introspective analysis that part was visual in origin and part auditory. Even in those cases in which, given conflicting optical and acoustic cues, I experienced two syllables, there was nothing about their quality that would have permitted me to know which I had seen and which I had heard.

By way of interpretation, MacDonald and McGurk (1978) indicate that their results bespeak a connection between perception and production, and McGurk and Buchanan (Note 4) echo a comment by Summerfield (1979), who observed, after having himself performed several experiments on the phenomenon, that the optical and acoustic signals are picked up in a "common metric of articulatory dynamics." I would agree, though I would, of course, prefer to call the common metric "phonetic." But a mode by any other name would bear as weightily on the issue I have put before you, for the important consideration is that, in any ordinary sense of modality, the speech percept is neither visual nor auditory; it is, rather, something else.

Integration into ordered strings. Having so far considered only the perception of individual phonetic segments, we should put some attention on the fact that phonetic segments are normally perceived in ordered strings. This wants explicit treatment if only because, as the reader may recall, a characteristic of the speech code is that several phonetic segments are conveyed simultaneously by a single segment of sound. As the reader may also recall, it is just this characteristic of the code that enables the listener to evade the limitation imposed by the temporal resolving power of the ear. The further consequence for perception, which we will consider now, is that the listener cannot perceive phonetic segment by phonetic segment in left to right (or right to left) fashion; rather, he must take account of the entire stretch of sound over which the information is distributed. Such an acoustic stretch typically signals a phonetic structure that comprises several segments. I will offer only a brief example, taken from a recent study by Repp et al. (1978), and chosen because the relevant span happens to cross a word boundary.

The experiment dealt with the effect of two cues, silence and noise duration, on perception of the locutions gray ship, gray chip, great ship, and great chip. In Figure 13 is a spectrogram of the words gray ship, with which the experiment began. The variable, shown in the figure, was the duration of silence between the two words. Given the results of previous research, we knew that increasing the silence would bias away from the fricative in ship and toward the affricate (stop-initiated fricative) in chip (Dorman, Raphael, & Isenberg, 1980; Dorman et al., 1979). The parameter, also shown in the figure, was the duration of the fricative noise, known from previous research to be a cue for the same distinction: increases in duration of the noise bias toward fricative and away from affricate (Gerstman, 1957; Dorman et al., 1979).

In Figure 14 are the results. We see in the graph at the upper left that when the noise duration was relatively short (62 msec), increasing the

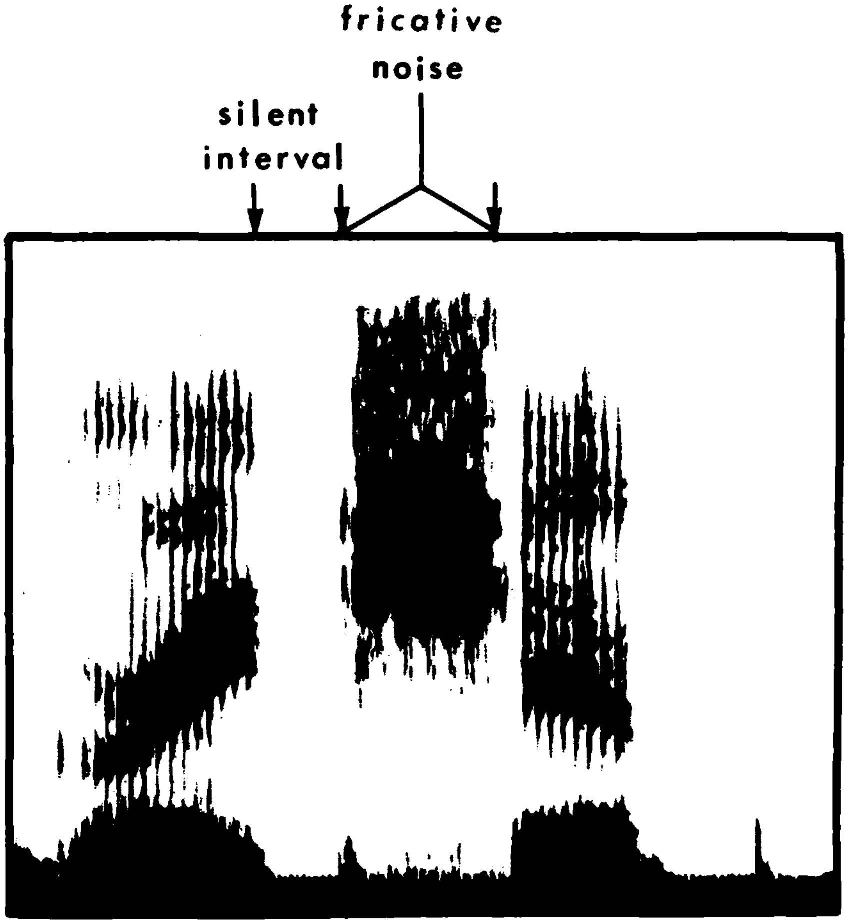


Figure 13. Spectrogram of the words "gray ship."

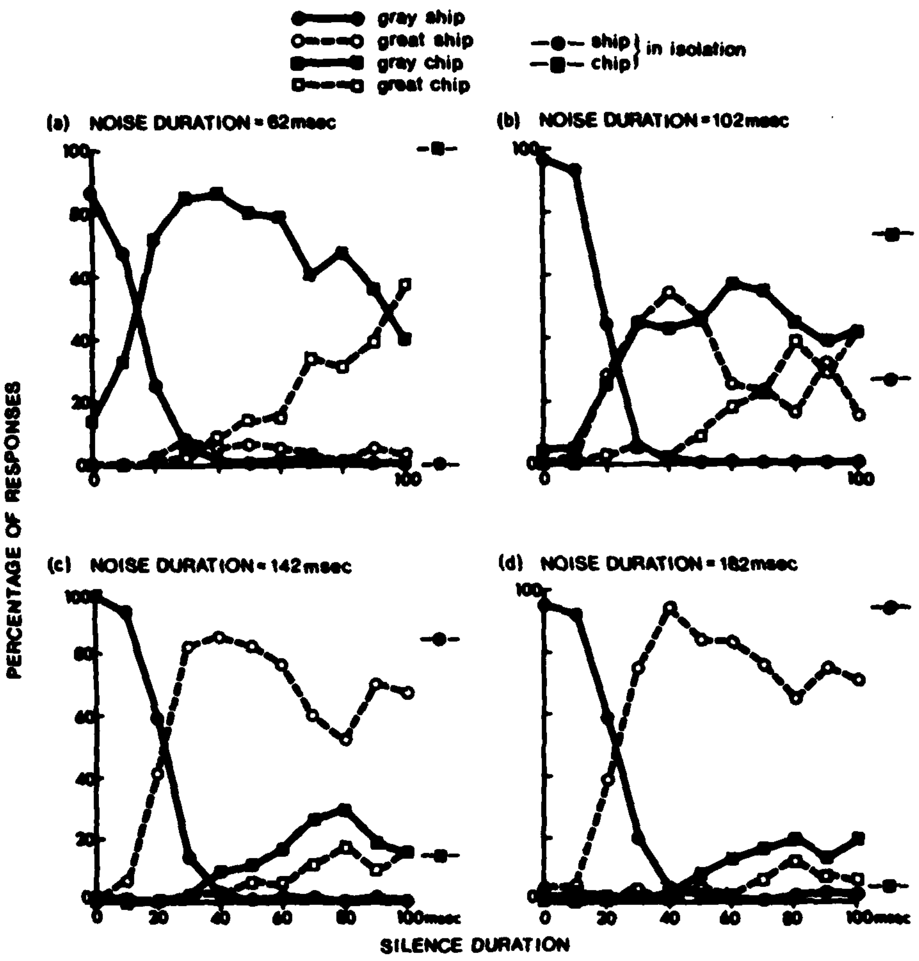


Figure 14. The effect of duration of silence, at each of four durations of fricative mode, on the perception and placement of stop (or affricate) manner. (From "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," by B. H. Repp, A. M. Liberman, T. Eccardt, and D. Pesetsky, Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637. Copyright 1978 by the American Psychological Association, Inc. Reprinted by permission.)

duration of the silence caused the percept to change from ship to chip. Thus, the effect of silence was to produce a stop-like consonant to its "right," much as it had done in the cases of slit-split and [sa]-[spa]-[sta] that were dealt with earlier. But, as shown in the graph at the lower right, when the duration of the fricative noise was relatively long (182 msec), increases in the duration of the silence caused the perception to change, not to gray chip, as before, but to great ship. That is, increasing the duration of the fricative noise in ship put a stop consonant at the end of the preceding word. The effect is superficially "right to left." But, of course, the effect is in neither direction; it is more properly regarded as a matter of apprehending a structure.

Given, then, that the listener must recover several phonetic segments from the same span of sound, we ask three questions about the underlying process. First, how does the listener delimit the acoustic span? That is, how does he know when all the information that is to be provided has been provided? There is, after all, no acoustic signal that regularly marks the information boundary. Second, how does the listener store the information as it accumulates? And, third, what does he do while he waits? Does he simply resonate, as it were, or does he entertain hypotheses? If the latter, does he entertain all possible hypotheses? Does he weight them according to the likelihood they are correct? And how quickly does he abandon them as they are proved wrong?

If these questions seem familiar to students of sentence perception, it is, I think, because processes in the phonetic and syntactic domains do have something in common. In both cases, information is distributed in distinctively linguistic ways through the signal. As a consequence, the perceiver must recover distinctively linguistic structures. To that extent, the resemblance between processing in the two domains is not superficial. Nor is it, if we take the vertical view of language I earlier espoused, altogether surprising.

Afterwords, Omissions, and Prospects

Having set out years ago to study communication by acoustic alphabets, we might still be so occupied. For acoustic alphabets can be used for communication--witness Morse code--and there are innumerable experiments we could have done had we gone on trying to find the alphabet that works best. But it is not likely, as a practical matter, that we would ever have made a large improvement. Nor is it likely, from a scientific point of view, that we would ever have learned anything interesting. Acoustic alphabets cannot become part of a coherent process; I suspect, therefore, that there is nothing interesting to be learned.

But speech was always before us, proof that there is a better way. Inevitably, then, we put our attention there and, in so doing, began to bark up the right tree. It remained only to find that speech and language require to be understood in their own terms, not by reference to diverse processes of a horizontal sort. But once the vertical view is adopted, there is little doubt about what we must try to understand.

There is also little doubt, at any stage of the research on speech, about how much or how little we do understand, because there is a standard by which progress can be measured; we are not in the position of explaining behavior that we have ourselves contrived. Thus, to test what we think we know of the relation between phonetic structure and sound, we have only to see how that knowledge fares when used as a basis for synthesis. In fact, it does well enough to enable us to synthesize reasonably intelligible speech, which suggests that we do know something (Liberman, Ingemann, Lisker, Delattre, & Cooper, 1959; Klatt, 1980; Mattingly, 1980). But the speech is not nearly so good as the real thing, which proves, as if proof were needed, that we have something still to learn. Perhaps what we must learn most generally is to accept the hypothesis, alluded to earlier in the paper, that human listeners are sensitive to all the phonetically relevant information in the speech signal. If that hypothesis is true, and if the acoustic cues that convey the information are as numerous, various, and intertwined as we now believe them to be, then we should act on our assumption that the key to the phonetic code is in the manner of its production. That requires taking account of all we can learn about the organization and control of articulatory movements. It also requires trying, by direct experiment, to find the perceptual consequences (for the listener) of various articulatory maneuvers (by the speaker). To do that we must, of course, press forward with the development of a research synthesizer designed to operate from articulatory, rather than acoustic, controls (Mermelstein, 1973; Rubin, Baer, & Mermelstein, 1981; Abramson, Nye, Henderson, & Marshall, 1981). The perfection of such a device, itself an achievement of some scientific consequence, will enable us to find a more accurate, elegant, and useful characterization of the informational basis for speech perception.

It will not have escaped notice that the claim to understanding I have made is, in any case, a modest one. At most, we presume to know something about what phonetic processes do, and in what ways they are distinctive and coherent. As for mechanism, however, there is only the assumed link between perception and production, and even there we have no certain, or even clear, idea how such a link might be effected. If we knew more about mechanism, we would presumably be in a better position to design automatic speech recognizers of a nontrivial sort (Levinson & Liberman, 1981). At present, however, we can only claim to understand where the difficulties lie. That is an important step, to be sure, but it is only the first one, and it will almost surely prove to be the easiest.

Since I have taken the position that speech perception depends on biologically specialized processes, I should, at last, acknowledge that neurological and developmental studies are relevant. For if phonetic processes are distinctive and coherent from a perceptual point of view, we reasonably expect that they are so from a neurological point of view as well. We do, then, look to neuropsychological data to provide further tests of our hypotheses, to refine our characterizations, and, indeed, to supply new insights into the processes themselves. As for the biology of the matter, we must rely heavily, of course, on developmental studies of speech perception, especially when these include very young infants and comparisons across languages. Such studies enlighten us about what might have developed by evolution in the history of the race, and what remains to develop, presumably by epigenesis, in the history of the individual. Of course, neither the neuropsychological nor the developmental studies will be useful unless we ask the right questions. But I believe we are learning how to do precisely that.

REFERENCE NOTES

1. Fodor, J. A. The modularity of mind. Unpublished manuscript, MIT, 1981.
2. Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. Integration of time-varying cues and the effects of phonetic context. Unpublished manuscript, Haskins Laboratories, 1981.
3. Lane, H. L., & Schneider, B. A. Discriminative control of concurrent responses by the intensity, duration and relative onset time of auditory stimuli. Unpublished report, Behavior Analysis Laboratory, University of Michigan, 1963.
4. McGurk, H., & Buchanan, L. Bimodal speech perception: Vision and hearing. Unpublished manuscript, Department of Psychology, University of Surrey, 1981.

REFERENCES

Abramson, A. S., Nye, P. W., Henderson, J. B., & Marshall, C. W. Vowel height and the perception of consonantal nasality. Journal of the Acoustical Society of America, 1981, 70, 329-339.

Bailey, P. J., & Summerfield, Q. Information in speech: Observations on the perception of [s]+stop clusters. Journal of Experimental Psychology: Human Perception and Performance, 1980, 6, 536-563.

Bailey, P. J., Summerfield, Q., & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. Haskins Laboratories Status Report on Speech Research, 1977, SR-51/52, 1-25.

Bastian, J., Delattre, P., & Liberman, A. M. Silent interval as a cue for the distinction between stops and semivowels in medial position. Journal of the Acoustical Society of America, 1959, 31, 1568. (Abstract)

Best, C. T., Morrongiello, B., & Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception & Psychophysics, 1981, 29, 191-211.

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. Perception & Psychophysics, 1981, 29, 26-36.

Cole, R. A., & Scott, B. Perception of temporal order in speech: The role of vowel transitions. Canadian Journal of Psychology, 1973, 27, 441-449.

Cooper, F. S. Research on reading machines for the blind. In P. A. Zahl (Ed.), Blindness: Modern approaches to the unseen environment. Princeton: Princeton University Press, 1950, 512-543.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 1952, 24, 597-606.

Cooper, F. S., Liberman, A. M., & Borst, J. M. The interconversion of audible and visible patterns as a basis for research on the perception of speech. Proceedings of the National Academy of Sciences, 1951, 37, 318-327.

Dorman, M. F. On the identification of sine-wave analogues of CV syllables. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), Proceedings of the Ninth International Congress of Phonetic Sciences (Vol. II). Copenhagen: University of Copenhagen, 1979, 453-460.

Dorman, M. F., Raphael, L. J., & Isenberg, D. Acoustic cues for fricative-affricate contrast in word-final position. *Journal of Phonetics*, 1980, *8*, 397-405.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, 1979, *65*, 1518-1532.

Fant, C. G. M. Descriptive analysis of the acoustic aspects of speech. *Logos*, 1962, *5*, 3-17.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, 1980, *27*, 343-350.

Gerstman, L. J. *Perceptual dimensions for the friction portions of certain speech sounds*. Unpublished doctoral dissertation, New York University, 1957.

Harris, K. S. Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1958, *1*, 1-7.

Harris, K. S., Hoffman, H. S., Liberman, A. M., & Delattre, P. C. Effect of third-formant transitions on the perception of the voiced consonants. *Journal of the Acoustical Society of America*, 1958, *30*, 122-126.

Isenberg, D., & Liberman, A. M. Speech and non-speech percepts from the same sound. *Journal of the Acoustical Society of America*, 1978, *64* (Suppl. 1), S20. (Abstract)

Klatt, D. H. Software for cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 1980, *67*, 971-995.

Levinson, S. E., & Liberman, M. Y. Speech recognition by computer. *Scientific American*, 1981, *244*, 64-76.

Liberman, A. M. The grammars of speech and language. *Cognitive Psychology*, 1970, *1*, 301-323.

Liberman, A. M. The specialization of the language hemisphere. In F. O. Schmitt & F. G. Worden (Eds.), *The Neurosciences: Third Study Program*. Cambridge, Mass.: MIT Press, 1974, 43-56.

Liberman, A. M. Duplex perception and integration of cues: Evidence that speech is different from nonspeech and similar to language. In E. Fischer-Jørgensen, J. Rischel, & N. Thorsen (Eds.), *Proceedings of the Ninth International Congress of Phonetic Sciences* (Vol. II). Copenhagen: University of Copenhagen, 1979, 468-473.

Liberman, A. M., & Cooper, F. S. In search of the acoustic cues. In A. Valdman (Ed.), *Papers in phonetics and linguistics to the memory of Pierre Delattre*. The Hague: Mouton, 1972, 329-338.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497-516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 1954, *68*, 1-13.

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. C., & Cooper, F. S. Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America*, 1959, *31*, 1490-1499.

Liberman, A. M., Isenberg, D., & Rakerd, B. Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, in press.

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held,

H. W. Leibowitz, & H.-L. Teuber (Eds.), Handbook of sensory physiology. Vol. VIII: Perception. New York: Springer-Verlag, 1978, 143-178.

Lisker, L. Rapid vs. rabid: A catalogue of acoustic features that may cue the distinction. Haskins Laboratories Status Report on Speech Research, 1978, SR-54, 127-132.

MacDonald, J., & McGurk, H. Visual influences on speech perception processes. Perception & Psycophysics, 1978, 24, 253-257.

Mann, V. A. Influence of preceding liquid on stop-consonant perception. Perception & Psychophysics, 1980, 28, 407-412.

Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. Further investigation into the influence of preceding liquids on stop consonant perception. Journal of the Acoustical Society of America, 1981, 69 (Suppl. 1), S91. (Abstract)

Mann, V. A., & Repp, B. H. Influence of vocalic context on perception of the [ʃ] -[s] distinction. Perception & Psychophysics, 1980, 28, 213-228.

Mann, V. A., & Repp, B. H. Influence of preceding fricative on stop consonant perception. Journal of the Acoustical Society of America, 1981, 69, 548-558.

Mattingly, I. G. Phonetic representation and speech synthesis by rule. Haskins Laboratories Status Report on Speech Research, 1980, SR-61, 15-21.

Mattingly, I. G., & Liberman, A. M. The speech code and the physiology of language. In K. N. Leibovic (Ed.), Information processing in the nervous system. New York: Springer Verlag, 1969, 97-114.

Mattingly, I. G., Liberman, A. M., Syrdal, A. M., & Halwes, T. Discrimination in speech and nonspeech modes. Cognitive Psychology, 1971, 2, 131-157.

McGurk, H., & MacDonald, J. Hearing lips and seeing voices. Nature, 1976, 264, 746-748.

Mermelstein, P. Articulatory model for the study of speech production. Journal of the Acoustical Society of America, 1973, 53, 1070-1082.

Nye, P. W. Psychological factors limiting the rate of acceptance of audio stimuli. In L. L. Clark (Ed.), Proceedings of the International Congress on Technology and Blindness. New York: American Foundation for the Blind, 1963, 99-109.

O'Connor, J. D., Gerstman, L. J., Liberman, A. M., Delattre, P. C., & Cooper, F. S. Acoustic cues for the perception of initial /w,j,r,l/ in English. Word, 1957, 13, 25-43.

Oden, G. C., & Massaro, D. W. Integration of featural information in speech perception. Psychological Review, 1978, 85, 172-191.

Rand, T. C. Dichotic release from masking for speech. Journal of the Acoustical Society of America, 1974, 55, 678-680.

Raphael, L. J., Dorman, M. F., & Liberman, A. M. Some ecological constraints on the perception of stops and affricatives. Journal of the Acoustical Society of America, 1976, 59 (Suppl. 1), S25. (Abstract)

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.

Repp, B. H. Perceptual integration and differentiation of spectral cues for intervocalic stop consosnnants. Perception & Psychophysics, 1978, 24, 471-485.

Repp, B. H. Accessing phonetic information during perceptual integration of temporally distributed cues. Journal of Phonetics, 1980, 8, 185-194.

Repp, B. H. Phonetic trading relationships and context effects: New experimental evidence for a speech mode of perception. Haskins Laboratories

Status Report on Speech Research, 1981, *SR-67/68*, this volume.

Repp, B. H. Two strategies in fricative discrimination. *Perception & Psychopysics*, in press.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 621-637.

Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulations. *Journal of the Acoustical Society of America*, 1981, *69*, 1154-1163.

Rubin, P., Baer, T., & Mermelstein, P. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 1981, *70*, 320-328.

Stevens, K. N., & Blumstein, S. E. The search for invariant acoustic correlates for phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981, 1-38.

Strange, W., Jenkins, J. J., & Edman, T. R. Identification of vowels in vowel-less syllables. *Journal of the Acoustical Society of America*, 1977, *61* (Suppl. 1), S39. (Abstract)

Studdert-Kennedy, M. Speech perception. *Language and Speech*, 1980, *23*, 45-66.

Studdert-Kennedy, M., & Cooper, F. S. High-performance reading machines for the blind: Psychological problems, technical problems and status. In R. Dufton (Ed.), *Proceedings of the International Conference on Sensory Devices for the Blind*. London: St. Dunstan's, 1966, 317-342.

Summerfield, Q. Use of visual information for phonetic perception. *Phonetica*, 1979, *36*, 314-331.

Summerfield, Q., Bailey, P. J., Seton, J., & Dorman, M. F. Fricative envelope parameters and silent intervals in distinguishing 'slit' and 'split.' *Phonetica*, 1981, *38*, 181-192.

Whalen, D. H., Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America*, 1981, *69*, 275-282.

FOOTNOTES

[1] At one point we assumed that these principles were so general as to extend to perception in all modalities. Indeed, we carried out experiments designed to explore the possibility that patterns could be preserved across vision and audition provided the stimulus coordinates were properly transformed (Cooper, Liberman, & Borst, 1951).

[2] In contrast to the remarkable sensitivity of the phonetic mode to all aspects of the acoustic signal that do convey phonetic information, there is its equally remarkable insensitivity to those aspects of the signal that do not. Thus, as is well known from many years of research on synthetic speech, the phonetic component of the percept is usually unaffected by gross variations in those aspects of the signal--for example, bandwidth of the formants--that are beyond the control of the articulatory apparatus and hence necessarily irrelevant for all linguistic purposes (Liberman & Cooper, 1972; Remez et al., 1981). The only effect of such variations is to make the speech sound unnatural or, in the most extreme cases, to make it impossible for the

listener to perceive the sound as speech.

[3] When the chirps are discriminated in isolation--that is, not as part of the duplex percept--the function has the same shape, but the level is displaced about 15 percentage points higher. The difference in level is presumably owing to the distraction produced in the duplex condition by the other side of the percept.

[4] The existence of these trading relations means that the location of a phonetic boundary on an acoustic continuum is not fixed; within limits it will move as the settings of the several cues are changed. The boundary will also move, of course, as a function of phonetic context. (See the discussion, above, of the effect of preceding context on the [da]-[ga] boundary and also, for example, Mann and Repp, 1981; Repp and Mann, 1981.)

READING, PROSODY, AND ORTHOGRAPHY

Deborah Wilkenfeld+

INTRODUCTION

Phonetic recoding effects in silent reading have been reported by a number of investigators employing a variety of experimental techniques (see review by Conrad, 1972) and testing in several languages and orthographic systems (Tzeng, Hung, & Wang, 1977; Erickson, Mattingly, & Turvey, 1977; Navon & Shimron, 1981). While the presence of a phonetic representation in reading has been convincingly demonstrated, the source of the effect and the role of the representation remain largely unexplored. The obvious explanation--that the effect results from a process of grapheme-to-phoneme conversion--is falsified by evidence for phonetic recoding in reading non-alphabetic orthographies (Tzeng et al., 1977; Erickson et al., 1977).

One strategy that might prove fruitful in untangling these puzzles is to specify what linguistic properties are embodied in the phonetic representation constructed by fluent readers. The presence of segmental phonetic features has been firmly established by the studies cited above, but evidence for suprasegmental features, such as word stress and sentence prosody, has not heretofore been sought, though readers' subjective reports suggest that these features are also present. Kleiman (1975) demonstrated an important role for phonetic recoding in the comprehension of written sentences, and since suprasegmentals have been shown to play a role in the perception of spoken utterances, evidence for suprasegmentals in the phonetic representation of written language--which itself marks only the grossest suprasegmental properties of sentences--would be tantalizing evidence for a model of reading based on a strong dependency of reading on speech perception.

In a small pilot experiment using the response bias technique (Mehler & Carey, 1967), the study reported here sought evidence that subjects encode word stress in silent reading on the level of the single word.

+Also University of Connecticut.

Acknowledgment. This work was supported by NICHD Training Grant HD 00321 to Ignatius G. Mattingly and Donald Shankweiler at the University of Connecticut. I am indebted to Joseph Kupin for his help with the statistical analysis of data and for the computer program used in the experiment; to Lyn Frazier for her suggestions in the early development of the method employed; and to Janet Fodor and Ignatius Mattingly for their patient support and many helpful discussions throughout the project.

STIMULI

Test items in this experiment were ten words chosen from among those English disyllabic homographs whose syntactic class depends on the placement of primary stress. For example, content is a noun when the first syllable is stressed and an adjective (or reflexive verb "to content oneself") when the second syllable is stressed. Similarly, permit is a noun when the first syllable is stressed and a verb when the second syllable is stressed. The orthography does not represent the location of word stress for these words; presumably in normal circumstances, sentential context provides the necessary information for choosing in these few ambiguous cases.

Test stimuli were lists composed of eight unambiguously stressed disyllabic words and a ninth, final word taken from the set of homographs. All of the unambiguous words in a single list were matched for placement of primary stress (i.e., all had first syllable stress or all had second syllable stress) but were of varied syntactic and semantic classes.

Test lists were embedded in a series of foil lists consisting of from eight to eleven words chosen at random. The ratio of foil sets to test sets was 7:1, yielding 80 lists.

In a pretest of the test stimuli, 20 subjects were asked to read aloud a list of 200 English words, among which the test words were embedded. Their assignment of stress for the homographs was recorded. Responses to this pretest were used as a baseline measure of preference in the experiment. Results appear in Table 1, Column A. Each test homograph was preceded in the main experiment by a list that shared the stress pattern of its less-preferred reading.

SUBJECTS

Subjects were 18 undergraduate volunteers enrolled in introductory linguistics courses at the University of Connecticut. All were native speakers of English. They were paid for their participation.

PROCEDURE

Subjects were told that the purpose of the main experiment was to measure the effect of reading rate on accuracy of recall. Each subject was tested separately. The subject was seated in front of a computer-controlled CRT screen on which appeared, for each trial, a vertical list of eight to eleven words. The subject was instructed to read each word on the list silently from top to bottom, as quickly as possible without missing any of the words, and to signal the experimenter when he or she was finished by reading the last word on the list out loud. The list on the screen then disappeared and was replaced by a single word. The subject was instructed to respond "yes" if the word was on the preceding list and "no" if it was not. This probe word was never one of the homographs. Subjects' spoken responses were tape-recorded for transcription later. The entire presentation took approximately fifteen minutes.

RESULTS

The results of this experiment are summarized in Table 1. Column A gives the percentage of times that the less-preferred stress pattern for each ambiguous item was given as a response in the pretest and notes whether the less-preferred reading was as a noun (with first syllable stress) or as a verb (with second syllable stress). Column B gives the percentage of the trials in which the less-preferred stress pattern was elicited in the biasing condition. The number of subjects is given in parentheses in this column. Column C gives the percentage of trials in which the less-preferred pattern was elicited from subjects who answered the word recognition question correctly for that test list. The number of subjects who answered correctly appears in parentheses.

Table 1

% Less Preferred Stress

	A	B	C
ITEM	PRETEST (N-20)	BIAS CONDITION	BIAS CONDITION (MEMORY QUESTION CORRECT)
conduct	10% (initial)	72% (18)	82% (11)
object	20 (final)	17 (18)	13 (15)
pervert	40 (initial)	77 (18)	77 (18)
present	30 (final)	28 (18)	29 (14)
digest	20 (initial)	39 (18)	38 (16)
progress	40 (final)	33 (18)	29 (14)
permit	20 (initial)	33 (18)	46 (11)
subject	30 (final)	33 (18)	33 (18)
incline	10 (initial)	0 (17)	0 (17)
project	30 (initial)	53 (17)	56 (17)

Comparison of Columns A and B indicates an effect of the biasing lists on the stress pattern of the ambiguous test items. In a Wilcoxon one-tailed test, this difference was significant at the .05 level.

The biasing effect becomes even more apparent if we take into account subjects' performance on the recognition test. Column C gives the results just for subjects who answered the memory question correctly for the list in question. Comparison of Columns A and C shows a significant difference at the .01 level.

A further indication that the biasing manipulation was responsible for the effect observed is that a strong correlation (r=.81) was found between performance on the recognition task and number of shifted responses, accounting for 66% of the variation between subjects. This correlation is graphed in Figure 1. The graph shows a wide range of subject performance. If we look at both ends of this range, at the two least successful and the two most successful subjects, we find that where performance on the memory task was 69-70 per cent, subjects gave the less-preferred reading only 20 per cent of the time, while the two subjects who answered 88 per cent of the recognition questions correctly gave the less-preferred reading 60 per cent of the time.

DISCUSSION

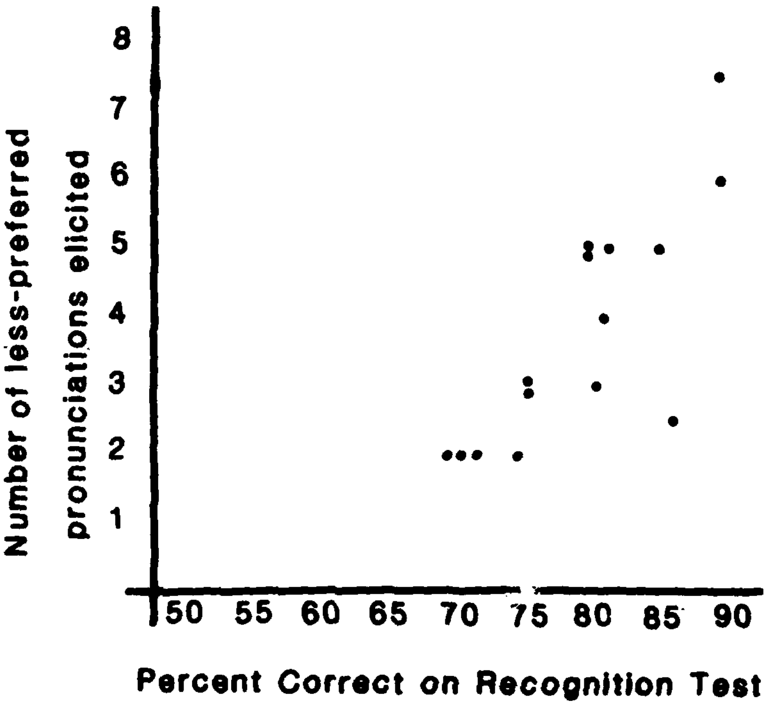
The correlation found is open to two interpretations. Under one interpretation, a subject's success in the recognition task is attributable to the amount of attention paid to the task. The more attentive subjects were more likely to have thoroughly read the word lists; thus they were more likely to have recoded the items on the list, and so to have been primed by properties of the code.

Under another more interesting interpretation, the more successful subjects did more phonetic recoding, as evidenced by the high likelihood that they would be primed by a phonetic property of the word lists. An _incidental_ result of this recoding was the ability to better remember what they had read, and thus better performance on the recognition test.

Under the first of these interpretations, attention rather than the requirements of the reading task _per se_ is what determines performance on the recognition test; the evidence found for mental representation of prosody is a by-product of a process, i.e., constructing the phonetic representation, which is perhaps just one of several representations constructed incidentally in the course of performing the experimental task.

Under the second interpretation, phonetic recoding is an integral part of good reading, and so if people are reading well, they must be constructing a phonetic representation. This will then prime pronunciation of the ambiguous item in the absence of contextual cues. The availability of the phonetic representation _incidentally_ facilitates performance on the recognition task. Better recognition results from greater ease of access to or more completeness of the phonetic representation, which may in turn indicate superior reading ability.

The first (attention) explanation suggests that any number of codes results from attending to the list, and does not give any reason to attribute special status to any code. Thus we should expect semantic and orthographic codes, for instance, to affect subjects' performance similarly to the phonetic code in memory tasks of the sort used in this experiment. The pattern of results reported for a similar task employed by Erickson et al. (1977),



Number of less-preferred pronunciations elicited
8
7
6
5
4
3
2
1
50
55
60
65
70
75
80
85
90
Percent Correct on Recognition Test

Figure 1

suggests that this is not the case; the orthographic and semantic properties of their word lists did not affect performance in a short-term recall task in the same way that the phonetic properties did.

It should be noted that the response shift was not equal for all the items tested. While a large effect was obtained for the words *digest*, *permit*, *project*, *conduct* and *pervert*, other items (*object*, *present*, *progress*) exhibited little effect (or even a reverse effect). *Incline* is the clearest case: in no trial was it possible to bias a subject in the test situation to pronounce *incline* as a verb, with second syllable stress. The averages given in Column A are for preferred pronunciations across twenty subjects. These figures indicate that one pronunciation of *incline*, for example, was preferred over the other by eighteen subjects out of twenty. What they do not indicate is how strong each individual's preference is. Though the former is much easier to measure, it provides only a very rough estimate of the latter--which is, of course, what is really relevant to the biasing experiment. The failure of the biasing manipulation for *incline* may well be due to the fact that while approximately one person out of ten prefers it as a noun, most people may have it in their lexicons only as a verb. For these people, its stress pattern would be completely unshiftable however psychologically real stress patterns are in reading. This suggests that for this kind of experiment it would be quite proper, and indeed optimal, to select words whose baseline frequency is about equal between noun and verb.

The objection might be made that the effect found in the present experiment is merely an artifact of the particular task employed, rather than a reflection of normal reading processes. To make this claim is to say that subjects employed strategies in the performance of this task that were constructed *ad hoc* for this purpose. But there is no logical requirement for such a strategy to include the construction of a phonetic representation; on the face of it, a visual representation would suffice. Nor is there any reason to expect all subjects to arrive at the same kind of special strategy. Yet the more successful subjects employed a phonetic coding strategy, while those subjects who could not do this did not seem to find another strategy that was similarly effective. Thus it appears that subjects were making the best use they could of reading skills that were already available for more ordinary purposes.

While it might be argued that the phonetic effects found by Conrad (1964) and Baddeley (1966), for example, and in the present experiment are due to rehearsal strategies for short-term recall, which have been shown to employ a phonetic representation (see Baddeley, 1976, Chapter 8, for discussion), this argument does not apply to effects found in the acceptability judgment task employed by Kleiman (1975), which did not require rehearsal. Thus the construction of a phonetic representation cannot be viewed as a mere artifact of rehearsal.

It could also be argued that for semantically integrated sentences, readers might use a semantic code, and employ a phonetic code to facilitate memory only when the items in the experimental sequence do not cohere semantically. The findings of Baddeley and Hitch (1974) address this criticism. They compared reaction times in a grammaticality judgment task using ordinary sentences and sentences composed of phonetically similar (rhyming) words. Phonetic similarity increased response latencies to grammatical and ungrammatical sentences. This task does not involve rehearsal or short-term memory. But it does implicate the parser, lending support to the conclusion from Kleiman's study that the sentence parsing mechanism requires a phonetic representation, quite apart from any requirements of short-term memory. If subjects construct a fairly detailed phonetic representation in a relatively unnatural situation in which it affords them no apparent advantage, we might also expect them to do it in a more natural situation. In other words, if subjects encode prosody when they read lists of words silently in a task that does not require comprehension, then it is likely that they will also encode prosody when they read ordinary sentences in a task that necessarily invokes the higher level processing involved in comprehension.

An important finding from this experiment is that readers construct a mental representation that includes features not represented in the stimulus. Thus, while it might be maintained that readers of English represent the segmental features of the words they read just because these can be extracted by rule from the letters of the orthographic system (at least in most cases), no such claim can be made for suprasegmental features such as stress, for there are no symbols in English orthography that indicate stress. In the stress-neutral pretest condition, subjects were always able to name the homographs. That this was not accomplished by simply applying rules to translate from orthography to phonology is strongly suggested by the fact that not all words having the same orthographic structure were consistently assigned the same pattern of stress by a single subject. More likely, a bias of some sort, due to factors such as frequency of occurrence, was responsible for a subject's choice in each case. Such a bias could only come from the lexicon. This is true in the case of vowel quality in homographs (lead, bow) as well. For these words, at least, naming written words must follow lexical access.

This must always be the case in naming Chinese logographs and Japanese kanji. These orthographic systems give very little phonological information, yet reading lists of words written in these orthographies results in a phonetic representation in short-term memory (Tzeng et al., 1977; Erickson et al., 1977). Thus almost all phonetic information must be supplied by the reader after lexical access.

Further support for the active participation of the lexicon in reading is provided by Hebrew. The Hebrew language is represented by an alphabetic orthography that keeps the vowel symbols fairly well separated from the consonant symbols. In texts intended for fluent adult readers, the vowel information is usually omitted entirely. However, it is the vowels of Hebrew that represent the inflectional system and carry most of the morphological and syntactic information. The task of the reader in Hebrew is to decide, presumably in the course of parsing procedures, the syntactic role of each word and its morphological composition in that role. Having derived this information, there is no reason to expect the reader of Hebrew to then add information about the vowels that would represent the word in speech. But the results of a study by Navon and Shimron (1981) suggest that they do indeed do so. Their subjects read lists of morphologically simple (uninflected) words in which vowel phonemes were represented by the optional vowel diacritics. Latencies in lexical decision tasks were increased by phonemically anomalous diacritics but not by graphemically anomalous diacritics that preserved the phonology. The effect could not be attributed to visual factors.

Their results suggest that in the simple case of reading unambiguous uninflected words, with no concurrent processing demands such as those required for sentence comprehension, subjects both construct a phonetic representation and access the lexicon. (In this case, lexical access appears to follow grapheme-to-phoneme translation. However, there is ample evidence, as Navon and Shimron point out, for models of lexical access that include a visual route. In any case, the result is a phonetic representation.) Yet Kleiman's results suggest that it is just in those cases in which processing for comprehension is required that the phonetic representation is important. In the case of fluent readers of Hebrew in the ordinary situation of reading text, the construction of a phonetic representation is at least as likely to occur as in the simple case of lexical decision. However, here the construction of the phonetic representation must follow lexical access, as with English homographs, Chinese logographs, and Japanese kanji. But with Hebrew, it is also likely to be the case that the phonetic representation is the product of the parser, rather than of the lexicon, since it is the analysis resulting from the parsing process that indicates to the reader what the morphology of the word must be, and thus what vowels must be supplied.

The facts about Hebrew, on the one hand, and English, Chinese and Japanese, on the other hand, suggest two hypotheses to account for the effect found in the present experiment. Under one hypothesis, which I will call the lexical bias hypothesis, prosodic priming is a result of activity in the lexicon. There is evidence that stress (or some abstract representation from which stress can be derived by rule [Chomsky & Halle, 1968]) is a feature of lexical entries (Brown & McNeill, 1966), just as segmental phonological features and semantic features are. As such, stress can probably be primed similarly to semantic features (Meyer, Schvaneveldt, & Ruddy, 1975). As the activation of a single word may activate any number of lexical entries in the same semantic field, the activation of a single disyllable with first-syllable stress might activate (if slightly) all disyllables having first-syllable stress. The activation of nine such words may have the cumulative effect of activating the first-syllable-stressed entry for the homograph to a point where it is much more readily available than the second-syllable-stressed entry, and thus more likely to be reported in the priming situation.

The second hypothesis, suggested by the facts about Hebrew, may be called the parsing hypothesis. According to this hypothesis, even isolated words are parsed, that is, they are processed as one-word sentences (see Mattingly, Note 1). It is in the parser that the morphophonemic representation retrieved from the lexicon is assigned a phonetic representation. This type of model is well suited to an orthography such as Hebrew. In fact, if it is assumed that the entire linguistic system, of which word recognition is only a part, is designed for the processing of linguistic structures, this type of model is equally well suited to English and any other language. The prosodic priming effect can then be seen as the result of a bias induced in the parser as it constructs a complete phonetic representation, including prosody, for each of a series of one-word sentences. A small bit of evidence in support of this hypothesis for English is the apparent ease with which sentences containing homographs are read: In syntactic context, the grapheme sequence p-r-o-g-r-e-

s-s (for example) may be instantly recognized as a noun or a verb as a result of information derived by the parser. The entire analysis of the sentence up to the point where the homograph is encountered determines what syntactic categories are likely to occur in a well-formed structure and guides lexical access to the appropriate entry, yielding, ultimately, the appropriate phonetic representation.

REFERENCE NOTE

1. Mattingly, I. G. *On the nature of phonological representations.* Manuscript in preparation, 1981.

REFERENCES

Baddeley, A. D. Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, 1966, *18*, 362-365.

Baddeley, A. D. *The psychology of memory*. New York: Basic Books, 1976.

Baddeley, A. D., & Hitch, G. Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8). New York: Academic Press, 1974.

Brown, R. W., & McNeill, D. The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 1966, *5*, 325-337.

Chomsky, N., & Halle, M. *The sound pattern of English*. New York: Harper & Row, 1968.

Conrad, R. Acoustic confusions in immediate memory. *British Journal of Psychology*, 1964, *55*, 75-84.

Conrad, R. Speech and reading. In J. Kavanagh & I. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading*. Cambridge, Mass.: MIT Press, 1972.

Erickson, D., Mattingly, I. G., & Turvey, M. T. Phonetic activity in reading: An experiment with Kanji. *Language and Speech*, 1977, *20*, 384-403.

Kleiman, G. M. Speech recoding in reading. *Journal of Verbal Learning and Verbal Behavior*, 1975, *14*, 323-339.

Mehler, J., & Carey, P. Role of surface and base structures in the perception of sentences. *Journal of Verbal Learning and Verbal Behavior*, 1967, *6*, 335-338.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. Loci of contextual effects on visual word-recognition. In P. M. A. Rabbitt (Ed.), *Attention and performance V*. London: Academic Press, 1975, 98-118.

Navon, D., & Shimron, J. Does word naming involve grapheme-to-phoneme translation? Evidence from Hebrew. *Journal of Verbal Learning and Verbal Behavior*, 1981, *20*, 97-109.

Tzeng, O. J. L., Hung, D. L., & Wang, W. S-Y. Speech recoding in reading Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory*, 1977, *3*, 621-630.

CHILDREN'S MEMORY FOR RECURRING LINGUISTIC AND NONLINGUISTIC MATERIAL IN RELATION TO READING ABILITY*

Isabelle Y. Liberman,+ Virginia A. Mann,++ Donald Shankweiler,+ and Michelle Werfelman+

Abstract. Good beginning readers typically surpass poor beginning readers in memory for linguistic material such as syllables, words, and sentences. Here we present evidence that this interaction between reading ability and memory performance does not extend to memory for nonlinguistic material like faces and nonsense designs. Using an adaptation of the continuous recognition memory paradigm of Kimura (1963) we assessed the ability of good and poor readers in the second grade to remember three different types of material: photographs of unfamiliar faces, nonsense designs, and printed nonsense syllables. For both faces and designs, the performance of the two reading groups was comparable; only when remembering the nonsense syllables did the good readers perform at a significantly superior level. These results support other evidence that distinctions between good and poor beginning readers do not turn on memory *per se*, but rather on memory for linguistic material. Thus they extend our previous finding that poor readers encounter specific difficulty with the use of linguistic coding in short-term memory.

The performance of good beginning readers on certain language-based short-term memory tasks, like their performance on many other language-related tasks, tends to be better than that of children who encounter difficulty in learning to read. The association between reading ability and such short-term memory skills is by now well-documented. For example, children who are good readers tend to have a better memory for strings of written or spoken letters (Shankweiler, Liberman, Mark, Fowler, & Fischer, 1979). They are also more

*In press, *Cortex*.

+Also University of Connecticut.

++Also Bryn Mawr College.

Acknowledgment. This work was partially supported by a grant from the National Institute of Child Health and Human Development (Grant HD01994) to Haskins Laboratories and by a NICHD postdoctoral fellowship HD05677) to the second author. For their generous cooperation on this project, we are also much indebted to the principals and second-grade teachers of the Northwest, Vinton and Southeast Elementary Schools in the Town of Mansfield, Connecticut. Special thanks are due to the 36 children who participated and to the parents who gave them permission to do so. In addition, we wish to express our appreciation to Leonard Mark who aided in testing the subjects, and to Robert Katz who provided assistance with the statistical analysis of the data.

[HASKINS LABORATORIES: Status Report on Speech Research SR-67/68 (1981)]

successful at recalling strings of spoken words, and even at recalling the words of spoken sentences (Mann, Liberman, & Shankweiler, 1980).

However, our concern has been not simply to document this performance difference but instead to uncover the probable cause of the difference. We first approached this problem by turning what appeared to us to be the special advantages of good readers against them. Since we knew that for adults, the presence of a high density of phonetically-confusable items hinders the use of speech-related processes in short-term memory, we were led to examine the effect of the same manipulation on the performance of good and poor readers. We found that like adults, good beginning readers appear to make effective use of phonetic coding in short-term memory, whereas poor readers do not. Thus we have shown that the memory performance of good readers falls sharply, even to the level of that of the poor readers, when they are asked to remember a letter string, word string, or sentence containing a high density of phonetically-confusable items (letters with rhyming names, or words that rhyme with one another), whereas the performance of poor readers remains little changed by this type of material.

At this point in our investigations, we were led to ask whether there are any other differences between the short-term memory capacities of good and poor readers, beyond those that reflect differential use of a speech code. After all, studies of patients with lateralized brain disease have revealed that verbal and non-verbal short-term memory abilities may be relatively independent (see, for example: Kimura, 1963; Milner & Taylor, 1972; Warrington & Shallice, 1969). Hence it seemed at least possible that the ability of poor readers to use nonverbal short-term memory processes could be equal to that of good readers. While this possibility is supported by findings that good and poor readers are equally successful at remembering unfamiliar (Hebrew) orthographic designs (Vellutino, Steger, Kaman, & DeSetto, 1975), it might seem inconsistent with findings that good readers surpass poor readers in remembering abstract figural patterns (Morrison, Giordani, & Nagy, 1977) and spatial-temporal patterns (Corkin, 1974). In our opinion, however, neither of these latter findings can be regarded as conclusive evidence that poor readers have difficulty with nonlinguistic short-term memory, _per se_, since both derive from materials that lend themselves to verbal labeling and to the use of linguistic memory strategies (Liberman, Mark, & Shankweiler, 1978). Therefore, it remained to be determined whether or not poor readers encounter difficulty with memory processes other than those requiring use of a speech code. We sought to investigate this question in the present study by comparing the ability of good and poor readers to remember linguistic material with their ability to remember material that is not only nonlinguistic but also not readily susceptible to linguistic coding.

Our subjects were good and poor readers in a second-grade classroom, whose memory abilities were tested with an adaptation of the continuous recognition memory paradigm of Kimura (1963). Using that paradigm, we assessed the children's ability to remember each of three types of materials: nonsense designs, photographs of unfamiliar faces, and printed CVC nonsense syllables. Whereas the nonsense designs were those employed in Kimura's original study (1963), the facial photographs and nonsense syllables were our own innovation. Studies of adult patients with focal brain damage reveal that the ability to encode and remember the nonsense designs that Kimura employed suffers as a consequence of right hemisphere temporal lobe excision but is relatively unimpaired by comparable excisions to the left, language-dominant, hemisphere (Kimura, 1963; Milner, 1974; Milner & Teuber, 1968). Likewise, the ability to encode and subsequently to recognize unfamiliar faces has been determined to be a right-hemisphere capacity that does not demonstrably depend on the language mediation skills of the left hemisphere (Leehey, Carey, Diamond, & Cahn, 1978; Yin, 1970). In contrast, the encoding and recognition of English-like nonsense syllables is a linguistic ability that suffers as a consequence of damage to the left hemisphere (Coltheart, 1980; Patterson & Marcel, 1977; Saffran & Marin, 1977).

We anticipated that the results obtained with good and poor readers in the case of nonsense designs and faces would differ from those obtained with nonsense syllables. Good readers were not expected to surpass poor readers in memory for either the nonsense designs or the faces, since neither of these sets of items lend themselves readily to the use of language coding. In the event, however, that good readers should excel at recognizing either of these materials, it would be taken as evidence that the poor readers do indeed have broader deficiencies in remembering. We expected good readers to surpass poor readers in memory for nonsense syllables, on the assumption that their use of phonetic coding as a mnemc.ic device would be superior to that of poor readers.

METHOD

Subjects

The subjects in this experiment were 36 second-grade children who attended the public schools in Mansfield, Connecticut. An initial pretest group was selected on the basis of the children's Total Reading Score on the Stanford Achievement Tests, which had been administered earlier in the same school year. Candidates for the good reading group had received grade scores of from 3.1 to 5.0, whereas candidates for the poor reading group had received scores of 1.5 to 2.4. Final selection of 18 good readers and 18 poor readers was made on the basis of scores on the Word Recognition Subtest of the Wide Range Achievement Test (WRAT) (Jastak, Bijou, & Jastak, 1965). Children selected as good readers had WRAT reading grade equivalents ranging from 3.1 to 5.0, with a mean score of 4.0; children selected for the poor reading group received grade equivalents from 1.5 to 2.4, with a mean score of 2.1.

Mean ages for good and poor readers were 94.0 months and 94.2 months, respectively, and were not significantly different. Individual administration of the WISC-R revealed good readers to have a mean Full Scale IQ of 113.6, with mean Verbal and Performance IQ's of 112.1 and 112.9, respectively. Poor readers received mean Full Scale IQ of 107.7, with Verbal and Performance IQ's of 104.9 and 109.1, respectively. There were no significant differences between good and poor readers on any of the IQ measures.

Materials

There were three different types of materials: nonsense designs, faces, and syllables. The tests using these three types of items were identical in manner of construction and presentation, each modeled on Kimura's (1963) recurring recognition memory task.

Nonsense designs. There were 80 nonsense-design stimuli, each of which was one of the 52 irregular line drawings of Kimura (1963). Four of the designs were used eight times each (the recurring designs), and the remaining 48 once each (the nonrecurring designs). Each stimulus was drawn on a 3 x 5 card. For the purpose of testing, the stimuli were divided into eight sets of ten; within each set of ten, the four recurring designs were randomly interspersed with six of the nonrecurring designs. The first set of ten stimuli constituted the inspection set, the remaining seven sets contained the actual test stimuli.

Faces. Face recognition stimuli were constructed using 52 black and white photographs, half of which were adult female faces and half adult male faces. In both the male and female stimuli sets, half were photographed looking to the left and half looking to the right. To minimize distinguishing details that might lend themselves to verbal labeling, no faces were used that displayed hair, eye-glasses, jewelry or distinctive markings such as scars, distinctive makeup, etc. In addition, a uniform mask was applied to each picture to cover hair and background detail as well as to ensure a uniform size.

Again, a set of 80 stimuli was constructed. Four photographs occurred eight times each (two male faces and two female faces, two looking to the left and two looking to the right) whereas the remaining 48 occurred once each. The stimuli were divided into eight sets each, with each set containing the four recurring photographs randomly interspersed among six nonrecurring ones. The first set served as the inspection set, the remaining seven sets contained the test stimuli.

Nonsense syllables. Stimuli for this part of the experiment were constructed from a set of 52 CVC nonsense syllables that had been selected from Hilgard (1962) to have a moderately low association value. Across the different syllables, frequency of occurrence of each letter was controlled as much as possible. The vowels *a*, *e*, and *u* appeared 11 times each, *i* appeared nine times, and *o* appeared ten times. Every consonant (with the exception of *q*, *x*, and *y* in initial position and *q*, *y*, *h*, and *w* in final position) occurred at least once, with some consonants occurring as often as six times.

From the syllables, a set of 80 stimuli was constructed. Four of the stimuli occurred eight times, while each of the remaining 48 occurred once. The stimulus cards were again divided into eight sets of ten each; within each set of ten the four recurring syllables were randomly interspersed with six non-recurring ones. The first set of ten constituted the presentation trials, the remaining seven sets contained the test stimuli.

Procedure

Each child was tested individually, with the nonsense designs being presented on the first day of testing, and the faces and syllables on a second day. The procedure for the recurring recognition memory paradigm was adapted from Kimura (1963) and was the same for all three types of material.

The experimenter began each test by telling the child that some designs (or faces or syllables), would be shown, one at a time, and that the task was to look at each one very carefully and try to remember it. She then presented the inspection set of ten cards, showing each card for approximately three seconds. Subsequently, the child was told that more cards would follow, some of which would be identical to those presented in the inspection set, and some of which would be new cards. The instruction was to say "Yes" if a card had been seen before, and "No" if it had not. The test items were then presented to the child, who was required to respond to each one before being shown the next.

RESULTS

In order to evaluate the performance of the subjects, we first computed the percentage of correct responses made by each subject, separately for each of the three types of materials (nonsense designs, faces, and syllables). This was done by summing the number of correct recognitions and correct rejections, and dividing by 70 (i.e., the total number of test items presented in each condition). After first noting that the performance of the subjects on all three types of material was consistently above the chance level of 50 percent correct, we turned to the major purpose of our study, which was to evaluate the extent of difference between the performance of good and poor readers on each of the three different types of items.

The results of an ANOVA computed on the variables of reading ability (good versus poor readers) and material type (designs, versus faces, versus syllables) revealed a significant effect of material type, F(2,68)=73.3, p<.001, reflecting the fact that designs and faces were typically harder to remember than syllables. There was further the anticipated interaction between the effect of item type and reading ability, F(2,68)=8.3, p<.001. As can be seen in Figure 1, good readers were not significantly better than poor readers at remembering either nonsense designs or faces. (For nonsense designs, t(34)=1.4, p>.1; for faces, t(34)=0.1, p>.6). In fact, poor readers were slightly (although not significantly) better at remembering nonsense designs. Good readers, however, were significantly better than poor readers at remembering the nonsense syllables, t(34)=3.2, p<.005.

DISCUSSION

The results, then, upheld our predictions. Poor readers were equal to good readers in ability to remember both nonsense designs and faces. In contrast, poor readers made significantly more errors than good readers in recognizing the nonsense syllables. Thus we find no evidence that children in the two reading groups differ in general memory ability. Rather, we again find them to differ only in memory for linguistic items. These findings help us to place in perspective two claims that are frequently made regarding the origins of many childhood reading problems. One claim sees a "general memory deficit" as central (Morrison et al., 1977). According to that hypothesis, which views poor readers as having difficulty with memory, _per se_, poor readers might be expected to show inferior performance for linguistic material and figural material alike. Clearly, our results are incompatible with this view, since it was found that good and poor readers differed solely in memory for the syllables.

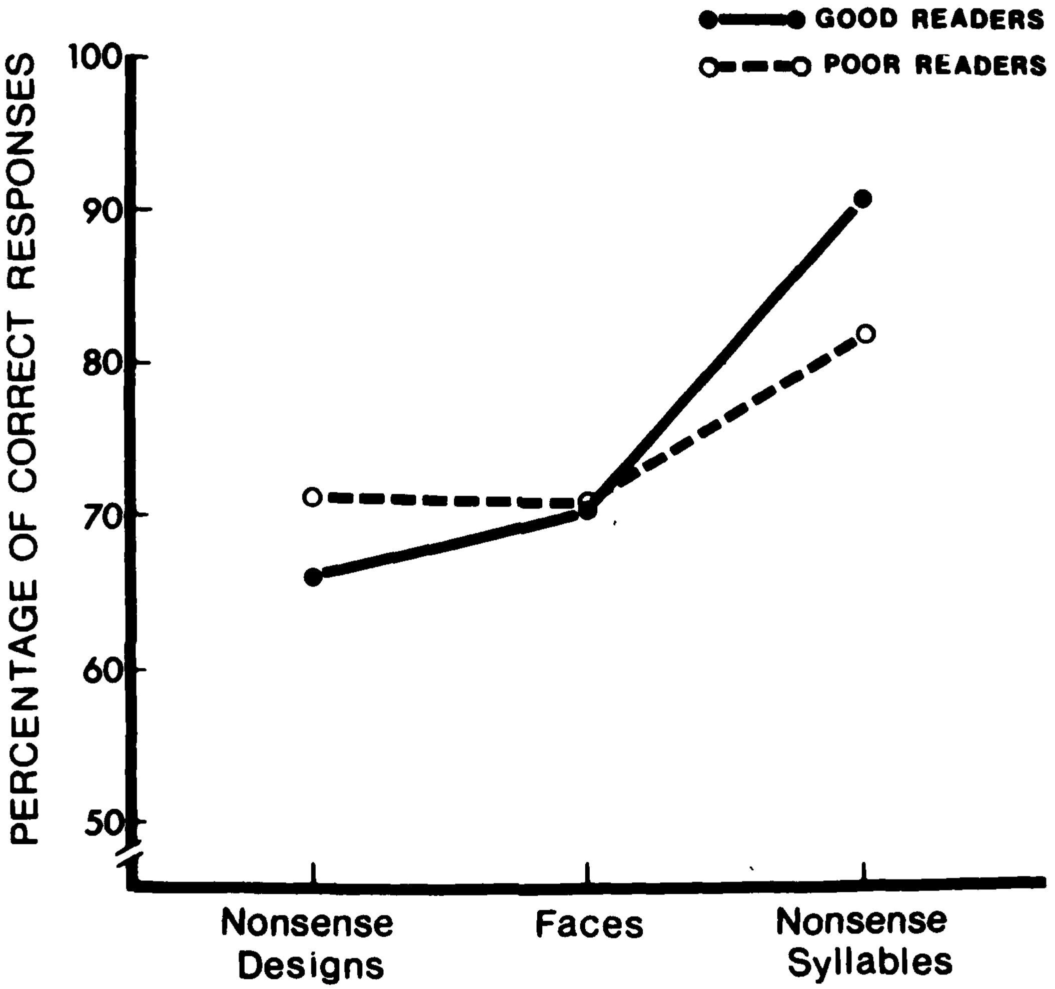


Figure 1: Mean percentage of correct responses made by good and poor readers on nonsense designs, faces, and nonsense syllables.

A second theoretical claim suggests that failure of serial order memory is the core problem (Bakker, 1972; Corkin, 1974; Holmes & McKeever, 1979). Our task did not require that subjects remember the order of items in the inspection set, yet we nonetheless obtained a difference between good and poor readers' ability to remember nonsense syllables. Thus the poor readers' memory problem goes beyond serial order alone. In this respect, the present findings confirm earlier results by Mark, Shankweiler, Liberman, and Fowler, 1977 and Byrne and Shea, 1979. We do realize, however, that a material-specific deficit in order memory could be a consequence of failure to make effective use of phonetic coding. Indeed, in a recent study (Katz, Shankweiler, & Liberman, Note 1) some of us found that good and poor readers selected by the same criteria as in the present study differed in ability to recall order of the items. But the good readers excelled only when their task was to recall the order of items that could be coded in terms of linguistic labels. No difference was found in memory for the order of nonrecodable items. Thus the problems of poor readers in recall of items, *per se*, and in recall of item order appears to be linked to some difficulty with using a phonetic code--either a failure to recode phonetically or a weakened tendency to use this coding principle.

In summary, then, we have discovered an instance in which despite identical procedures, good and poor readers differ in the ability to remember language-based material, but fail to differ in memory for two types of nonverbal material. Thus we conclude that the short-term memory deficits of poor readers appear indeed to be restricted to the domain of phonetic representation in short-term memory. Several questions arise at this point, among them the question of why poor readers fail to make effective use of a phonetic code, and the question of how a deficient linguistic memory comes to be associated with problems in learning to read. At present we are addressing the first of these questions by examining the pattern of memory errors made by poor readers. Our approach to the second, however, is guided by a consideration of the relation between short-term memory and normal language processing (Baddeley, 1978; Liberman, Mattingly, & Turvey, 1972), which leads us to ask whether poor readers encounter difficulty on the type of language comprehension tasks used in studying aphasic patients (Caramazza & Zurif, 1978). We suspect that answers to these two questions may bring us closer to an understanding of the reading process as well as of the process of reading acquisition.

REFERENCE NOTE

1. Katz, R., Shankweiler, D., & Liberman, I. Y. *Memory for item order and phonetic coding in the beginning reader*. Manuscript submitted for publication, 1981.

REFERENCES

Bakker, D. J. *Temporal order in disturbed reading*. Rotterdam: Rotterdam University Press, 1972.

Baddeley, A. D. The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. *Psychological Review*, 1978, *85*, 139-152.

Byrne, B., & Shea, P. Semantic and phonetic memory in beginning readers.

Memory & Cognition, 1979, 7, 333-341.

Caramazza, A., & Zurif, E. B. The comprehension of complex sentences in children and aphasics: A test of the regression hypothesis. In A. Caramazza & E. B. Zurif (Eds.), Language acquisition and language breakdown: Parallels and divergences. Baltimore: Johns Hopkins University Press, 1978.

Coltheart, M. Deep dyslexia: A right hemisphere hypothesis. In M. Coltheart, K. Patterson, & J. Marshall (Eds.), Deep dyslexia. Boston: Routledge & Kegan-Paul, 1980.

Corkin, S. Serial-order deficits in inferior readers. Neuropsychologia, 1974, 12, 347-354.

Hilgard, E. R. Methods and procedure in the study of reading. In S. S. Stevens (Eds.), Handbook of experimental psychology. New York: Wiley, 1962.

Holmes, D. R., & McKeever, W. F. Material specific serial memory deficit in adolescent dyslexics. Cortex, 1979, 15, 51-62.

Jastak, J., Bijou, S. U., & Jastak, S. R. Wide Range Achievement Test. Wilmington, Del.: Guidance Associates, 1965.

Kimura, D. Right temporal-lobe damage. Archives of Neurology, 1963, 8, 264-271.

Leehey, S., Carey, S., Diamond, R., & Cahn, A. Upright and inverted faces: The right hemisphere knows the difference. Cortex, 1978, 14, 441-449.

Liberman, A. M., Mattingly, I. G., & Turvey, M. Language codes and memory codes. In A. W. Melton & E. Martin (Eds.), Coding processes in human memory. Washington, D.C.: Winston, 1972.

Liberman, I. Y., Mark, L. S., & Shankweiler, D. Reading disability: Methodological problems in information-processing analysis. Science, 1978, 200, 801-802.

Mann, V. A., Liberman, I. Y., & Shankweiler, D. Children's memory for sentences and word strings in relation to reading ability. Memory & Cognition, 1980, 8, 329-335.

Mark, L. S., Shankweiler, D., Liberman, I. Y., & Fowler, C. A. Phonetic recoding and reading difficulty in beginning readers. Memory & Cognition, 1977, 5, 623-629.

Milner, B. Hemispheric specialization: Scopes and limits. In F. O. Schmitt & F. G. Worden (Eds.), The neurosciences: Third study program. Cambridge, Mass.: MIT Press, 1974.

Milner, B., & Taylor, L. Right hemisphere superiority in tactile pattern recognition after cerebral commissurotomy: Evidence for nonverbal memory. Neuropsychologia, 1972, 10, 1-15.

Milner, B., & Teuber, H-L. Alteration of perception and memory in man: Reflections on methods. In L. Weiskrantz (Ed.), Analysis of behavioral change. New York: Harper and Rowe, 1968.

Morrison, F. J., Giordani, B., & Nagy, J. Reading disability: An information-processing analysis. Science, 1977, 196, 77-79.

Patterson, K. E., & Marcel, A. J. Aphasia, dyslexia and the phonological coding of printed words. Quarterly Journal of Experimental Psychology, 1977, 29, 307-318.

Saffran, E. M., & Marin, O. S. M. Reading without phonology: Evidence from aphasia. Quarterly Journal of Experimental Psychology, 1977, 29, 515-525.

Shankweiler, D., Liberman, I. Y., Mark, L. S., Fowler, C. A., & Fischer, F. W. The speech code and learning to read. Journal of Experimental

Psychology: Human Learning and Memory, 1979, *5*, 531-545.

Vellutino, F. R., Steger, J. A., Kaman, M., & DeSetto, L. Visual form perception in deficient and normal readers as a function of age and orthographic familiarity. *Cortex*, 1975, *11*, 22-30.

Warrington, E. K., & Shallice, T. The selective impairment of auditory-verbal short-term memory. *Brain*, 1969, *92*, 885-896.

Yin, R. K. Face recognition by brain-damaged patients: A dissociable disability? *Neuropsychologia*, 1970, *8*, 395-402.

PHONETIC AND AUDITORY TRADING RELATIONS BETWEEN ACOUSTIC CUES IN SPEECH PERCEPTION: PRELIMINARY RESULTS

Bruno H. Repp

Abstract. When two different acoustic cues contribute to the perception of a phonetic distinction, a trading relation between the cues can be demonstrated if the speech stimuli are phonetically ambiguous. Do the cues trade also in unambiguous stimuli? Four different trading relations were examined using a fixed-standard AX discrimination task with stimuli either from the vicinity of the phonetic category boundary or from within a phonetic category. The results suggest that certain trading relations (presumably of auditory origin) hold in both conditions while others are tied to the perception of phonetic contrasts and thus appear to be specific to the speech mode.

INTRODUCTION

Virtually any phonetic distinction has multiple correlates in the acoustic speech signal. That is, the articulatory adjustments required to change from one phonetic category to the other (other things equal) cause acoustic changes along several separable physical dimensions--spectrum, amplitude, time. While a listener typically perceives only a single change--viz., one of phonetic category--the physical changes that led to this unitary percept can only be described in the form of a list with multiple entries. When the signal properties thus listed are manipulated individually in an experiment, it is generally found that they all have perceptual cue value for the relevant phonetic distinction, although tney may differ in their relative importance. If one cue in such an ensemble is changed to favor category B, another cue can be modified to favor category A, so that the phonetic percept remains unchanged. This is called a trading relation. Presumably, any two cues for the same phonetic distinction can be traded off against each other within limits set by their acceptable range of values and by their relative perceptual weights. Numerous recent studies of trading relations have been reviewed by Repp (1981b); some of them will be discussed further below.

The mechanisms by which a listener's brain combines a number of diverse cues into a single phonetic percept are not known, but there are two

Acknowledgment. This research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to Haskins Laboratories. I am grateful to Janette Henderson for assistance in running subjects, and to Robert Crowder, Virginia Mann, Richard Pastore, and Sigfrid Soli for helpful comments on an earlier draft of the paper.

contrasting views on that issue. One view (e.g., Liberman & Studdert-Kennedy, 1978; Repp, Liberman, Eccardt, & Pesetsky, 1978) holds that the perceptual integration of acoustic cues is motivated by their common origin in the production of a phonetic contrast; that is, listeners are assumed to possess and apply detailed tacit knowledge of the multiple acoustic correlates of articulatory maneuvers. The other view (best spelled out in Pastore, 1981) maintains that integration of, and trading relations between, acoustic cues might arise either from integration or from interactions (such as masking or contrast) at a purely auditory level of processing, without reference to the articulatory origin of the cues. The evidence so far (summarized in Repp, 1981b) strongly favors the first view. However, it is conceivable that, as more is learned about auditory mechanisms, certain trading relations between acoustic cues will find auditory explanations, particularly those that seem to have no good articulatory rationale. Since many perceptual trading relations have been demonstrated with synthetic stimuli and without a parallel examination of speech production, the relation of the perceptual results to what happens in articulation may not always be as close as has been supposed, and some trading relations may actually have been caused by auditory cue interactions.

Undoubtedly, detailed studies of speech production and speech acoustics as well as auditory psychophysics will shed further light on this issue. There is a more direct experimental approach, however, which makes use of the fact that, under certain circumstances, the same (or highly similar) stimuli may be heard either as speech or as nonspeech. Such different percepts may be achieved either by presenting speechlike stimuli to human listeners under different instructions, relying primarily on the subjects' postexperimental reports about whether the stimuli in fact sounded speechlike or not, or by contrasting human perception of speech with that of nonhuman animals. In either case, the demonstration of a trading relation in all subjects or in all conditions would favor an auditory account, while the finding that a trading relation holds only when human listeners claim to perceive the stimuli as speech, but not when they claim to hear nonspeech sounds or when the listeners are nonhuman, would constitute strong evidence in favor of the speech-specific (articulatory-phonetic) origin of the trading relation.

There are no completed studies of trading relations in animals, but interesting results are expected soon from several laboratories. For chinchillas, Kuhl and Miller (1978) have reported a shift in the voicing boundary for stop consonants with place of articulation--an effect that may, in part, be due to a trading relation between voice onset time and formant onset frequencies (cf. Summerfield & Haggard, 1977). A trading relation between these two variables has also been demonstrated in human infants (Miller & Eimas, Note 1); however, rather than pointing towards psychoacoustic interactions, this finding may indicate that human infants are biologically prepared for phonetic perception. The present experiments focus on several effects that have not yet been demonstrated in either infant or animal subjects.

In studies using adult human subjects, two methods have been applied to address the question of the origin of trading relations. One is to construct stimuli that contain the critical cues under investigation but are sufficiently different from speech in other respects, so as to be perceived as nonspeech by naive subjects but as speech by more experienced or specially instructed subjects. The technique of imitating the speech formants with pure tones has served this purpose well (Bailey, Summerfield, & Dorman, 1977; Best, Morrongiello, & Robson, 1981; Remez, Rubin, Pisoni, & Carrell, 1981). The other method is to use speech stimuli and to lead listeners, through special instructions and practice, to perceive them analytically—to segregate them into their auditory components, as it were. This is a notoriously difficult task, but it is possible with certain special stimuli, e.g., with fricative-vowel syllables (Repp, 1981a). In all of these studies--some of which will be described in more detail below—subjects' response patterns were radically different when the stimuli were heard as speech than when the same stimuli were heard as nonspeech; in particular, the trading relations or other contextual effects under investigation were observed only in the speech mode. However, as noted above, this result may not hold for all trading relations.

The present experiments explored a third method, which has the advantage of simplicity and general applicability, thus making possible the parallel investigation of a number of different trading relations. The method is a simplified version of a procedure used by Fitch, Halwes, Erickson, and Liberman (1980) to demonstrate the categorical perception of speech stimuli varying in two cue dimensions. Fitch et al. were concerned with a trading relation between a temporal and a spectral cue for the "slit"-"split" contrast: the amount of silence between the fricative noise and the periodic stimulus portion, and the presence or absence of formant transitions (appropriate for a labial stop) at the onset of the periodic portion. In an identification task, less silence was needed to change "slit" to "split" when formant transitions were present than when they were absent. In a subsequent oddity discrimination task, Fitch et al. compared performance on three types of trials: (1) Spectral difference only ("one-cue condition"); (2) spectral and temporal difference, the stimulus with the formant transitions always having the longer silence ("two-cooperating-cues condition"); and (3) spectral and temporal difference, but the stimulus with the formant transitions now having the shorter silence ("two-conflicting-cues condition"). Subjects were considerably more accurate in the second than in the third condition, with performance in the first condition in between. This ordering of conditions was predicted from the way the stimuli were labeled by the subjects. In essence, these results revealed that speech stimuli varying on two dimensions are still categorically perceived. The listeners appeared to base their discrimination judgments on the phonetic labels of the stimuli, and thus the trading relation between the two cues was exhibited in discrimination as well as in labeling responses.

What would happen, however, if subjects could not rely on phonetic labels? Such a situation would arise if the stimuli to be discriminated were perceived as belonging to the same phonetic category. We know from many earlier studies of categorical perception that such discriminations are difficult to make, but subjects typically perform at a level better than chance and their performance may be enhanced by increasing physical stimulus differences and/or by using a paradigm that reduces stimulus uncertainty. If subjects cannot rely on phonetic labels, they must make their discriminations on the basis of the auditory properties of the stimuli. If some of these properties interact at the auditory level of perception and thereby generate a trading relation, then this trading relation should be observed regardless of whether or not listeners can make phonetic distinctions. On the other hand, if a trading relation is phonetic in origin, then the unavailability of phonetic contrasts should lead to a disappearance of the trading relation. Since, in this case, the cues are presumably independent at the auditory level, a difference in two cues should be at least as easy to discriminate as a difference in one cue (cf. Espinoza-Varas, Note 2), regardless of whether the cue values are paired in the cooperating or the conflicting manner (a la Fitch et al., 1980).

This is the rationale underlying the present experiments. To simplify the design, the cooperating-cues condition was omitted. The critical comparison was between 1-cue and 2-cue (conflicting-cues) trials in two discrimination conditions: Between phonetic categories and Within a single phonetic category. A trading relation in the Between condition (where stimuli contrasted phonetically on some, but not all, trials) should show up as poorer performance on 2-cue than on 1-cue trials. The same pattern in the Within condition would suggest that the trading relation is auditory in origin. On the other hand, equal or better performance on 2-cue than on 1-cue trials in the Within condi'.ion would indicate that the trading relation is absent and, therefore, that its occurrence in the Between condition has a phonetic basis.

Four different trading relations were investigated in four parallel experiments that were identical except for the stimuli and their dimensions of variation. Therefore, the general method will be described first, followed by a discussion of the individual experiments.

GENERAL METHOD

Stimulus Tapes

Each experiment employed speech stimuli (natural or synthetic words) varying on two cue dimensions for a specific phonetic contrast. One cue--the primary cue--was always temporal in nature and assumed several different values, whereas the other cue--the secondary cue--assumed only two different values. Two sets of four values of the primary cue were selected: One set of shorter values was intended to span the phonetic category boundary (Between condition), while the other set had longer values intended to fall entirely within the corresponding phonetic category (Within condition). Because Weber's Law holds approximately for the discrimination of duration (e.g., Creelman, 1962), and to facilitate discrimination in the more difficult Within condition, the values in the Within stimulus set were spaced farther apart than those in the Between set. The two values of the secondary cue were chosen so as to be difficult to discriminate but still sufficiently different to generate an observable trading relation.

A fixed-standard AX (same-different) discrimination task was used. This task has several advantages, which include low stimulus uncertainty (which tends to raise discrimination scores), relatively short test duration, and direct convertibility of the data into d' scores. The stimulus tapes for the Between and Within conditions were identical except for the settings of the primary cue. The fixed standard stimulus occurred first in each stimulus pair and was constant throughout each condition; it had the shortest value of the primary cue and the more conflicting of the two values of the secondary cue (i.e., that value which, more than the other value, favored the same phonetic category as did an increase in primary cue duration). Each condition contained four blocks of stimulus pairs. The first block of 48 pairs was for practice only: On half the trials, the standard was paired with itself; on the other half, it was followed by that stimulus which had the longest value of the primary cue but the same value as the standard of the secondary cue. In other words, the practice block contained only identical and (relatively easy) 1-cue trials. The first test block of 72 pairs contained the same pairs as the practice block plus 24 2-cue trials. On these latter trials, the difference in the primary cue between the standard and comparison stimuli was the same as on 1-cue trials, but there was an added difference in the secondary cue whose setting in the comparison stimulus "conflicted" with its longer value of the primary cue, thus making discrimination more difficult if (and only if) the two cues engaged in the predicted trading relation. The remaining two test blocks of 72 trials each were similar except that the magnitude of the difference in the primary cue was reduced, thus making the task increasingly more difficult. This was done to counteract possible ceiling effects due to individual differences in discrimination accuracy and in phonetic boundary locations. It also served to explore a range of stimulus differences, since it was not known in advance how well naive subjects would perform in this task.

The standard and comparison stimuli in a pair were separated by 500 msec of silence. The interpair interval was 2 sec, and there were longer pauses between blocks.

Procedure

The subjects were tested individually or in small groups. The stimuli were presented over TDH-39 earphones at a comfortable intensity. All subjects listened first to the Within condition, followed by the Between condition and by a repetition of the Within condition. The repetition served to investigate whether experience with phonetic contrasts in the Between condition had any effect on subjects' strategies in the Within condition; it also gave a second chance to those subjects who found this condition very difficult the first time. In all experiments except the first, the discrimination tests were followed by a brief labeling test in which the seven different stimuli used in the Between condition were presented 10 times in random order. (The labeling test for Exp. 1 was administered at the end of Exp. 4b.) This test was added to verify the trading relation between the two cues.

Instructions were kept to a minimum. The subjects were told about the general procedure and about the relative difficulty of the task. They were not informed about the difference between the two experimental conditions (except that the stimuli would be slightly different), and they were not told the relevant phonetic labels or the auditory cue dimensions that varied. Rather, they were left to discover these by themselves as they listened to the 48 practice trials. For these trials only, the correct responses (s, d) were printed on the answer sheet, and the subjects merely checked them off as they went along. It was hoped that, after this experience, the subjects would have some idea of the difference to listen for (i.e., that in the primary cue dimension). They were told that the differences in the subsequent test blocks were of the same kind, but that they would get smaller in magnitude. They were not informed about the introduction of another kind of difference (that in the secondary cue dimension) or of the consequent increase in the true proportion of "different" trials from 50 to 67 percent, but it was mentioned that any kind of difference perceived warranted a response of "different." Clearly, the procedure was designed to focus the subjects' attention on the primary cue, since only this cue varied in the practice block.

The subjects responded by writing down "s" or "d" on each trial, guessing if necessary. After each of the three test conditions, they were interviewed about their impressions and strategies. In the final labeling test, they chose from the two relevant categories (which they were told) and wrote down their responses in abbreviated form.

Analysis

Individual subject scores in each test block were converted into d' values, taking the proportions of "different" responses on 1-cue and 2-cue trials, respectively, as separate hit rates, and the proportion of "different" responses on trials of identical stimuli as the joint false-alarm rate. Proportions of 0 and 1 were treated as .01 and .99, respectively, thus limiting d' to a maximum value of 4.66.

Three analyses of variance were conducted on subjects' d' scores in each experiment. The first analysis was on the Between condition only, with the factors Cues (1-cue vs. 2-cue) and Blocks (three levels of difficulty). The second analysis was on the Within condition only, with the factors Repetitions, Cues, and Blocks. (In Exp. 3, only the second repetition was analyzed.) The absence of any interactions between Repetitions and the other factors justified the combination of the two repetitions for the third analysis, which compared the Between and Within conditions with the factors Conditions, Cues, and Blocks. The critical effect in this last analysis was the Conditions by Cues interaction, which was expected to reveal whether or not the same trading relation (or other response pattern) held in the two conditions.

EXPERIMENT 1: "SAY"-"STAY"

The trading relation studied here concerned, as the primary cue, the amount of silence following the fricative noise and, as the secondary cue, the onset frequency of the first formant (F1) following the silence. This trading relation, which is similar to that for "slit"-"split" studied by Fitch et al. (1980), has been previously investigated by Best et al. (1981): Less silence is needed to change "say" to "stay" when F1 starts at a lower frequency. Best et al. confirmed this trading relation in two different discrimination tests (oddity and variable-standard AX). These tests actually included some within-category trials along with between-category trials, and the trading relation could be seen to disappear within the "stay" category. However, this result is not conclusive, since it may reflect a floor effect and is based on rather few responses. It is interesting to note, however, that the similar data of Fitch et al. (1980) for the "slit"-"split" contrast, although they are open to the same objections, actually suggest a reversal of the trading relation in the within-category regions: Whereas the ordering of performance on the three types of trials was cooperating cues > one cue > conflicting cues in the phonetic boundary region, it changed to cooperating cues = conflicting cues > one cue (at chance) within categories. This is exactly the pattern one should expect from a trading relation that is specific to phonetic perception.

This expectation was further confirmed by Best et al. (1981) in an elegant study with "sine-wave analogs" of "say"-"stay" stimuli. Subjects who reported that they heard the sine-wave stimuli as (highly unnatural) tokens of "say" or "stay" exhibited the same trading relation between silence duration and F1(-analog) onset frequency as was observed in speech stimuli, whereas those subjects who heard the sine-wave stimuli as nonspeech showed a radically different pattern of responses that suggested that they paid selective attention to variations in one or the other cue. They neither integrated the cues into a unitary percept, nor did the settings of the unattended cue have much effect on the perception of the attended cue.

Given these rather convincing results, the present re-investigation of the "say"-"stay" contrast served not only to replicate the findings of Best et al. but also to validate the new procedure. The prediction was, then, that the trading relation between silence duration and F1 onset frequency would be observed only in the Between condition but not in the Within condition.

Method

Subjects. Eleven volunteers were recruited by announcements on the Yale University campus and were paid for their participation. Most of them had served in earlier speech perception experiments. A different group of 9 subjects (those of Exp. 4b) took the brief labeling test.

Stimuli. The stimuli were hybrids composed of a natural-speech [s] noise followed by a synthetic periodic portion. The [s] noise derived from a male speaker's utterance of [sa]. The periodic portion was produced on the OVE IIIc serial resonance synthesizer at Haskins Laboratories, following formant specifications provided by Best et al. (1981) in their Figure 1 (speaker SSB). The fricative noise was 212 msec long, with a gradually rising amplitude over the first 170 msec and a rapid fall thereafter. The duration of the synthetic periodic portion was 300 msec. It had a fairly abrupt onset and a fundamental frequency that fell linearly from 110 to 80 Hz.

The two stimulus portions were concatenated after both had been digitized at 10 kHz using the Haskins Laboratories PCM system. The primary cue was the amount of silence between them. In the Between condition, the standard stimulus had no silence at all ("say"), and the comparison stimuli had 30, 20, and 10 msec, respectively, on "different" trials in the three test blocks. In the Within condition, the standard had 70 msec of silence ("stay"), and the comparison values were 130, 115, and 100 msec. The "say"-"stay" boundary was expected to be in the vicinity of 20 msec of silence. The secondary cue was the onset frequency of F1 in the periodic portion. On 1-cue trials, it was 200 Hz, whereas, on 2-cue trials, it was raised to 299 Hz—a cue favoring "say" and thus "conflicting" with the longer silence cue in the comparison stimuli. The difference in F1 between the two versions of the periodic stimulus portions gradually diminished over the first 40 msec (the extent of the F1 transition).

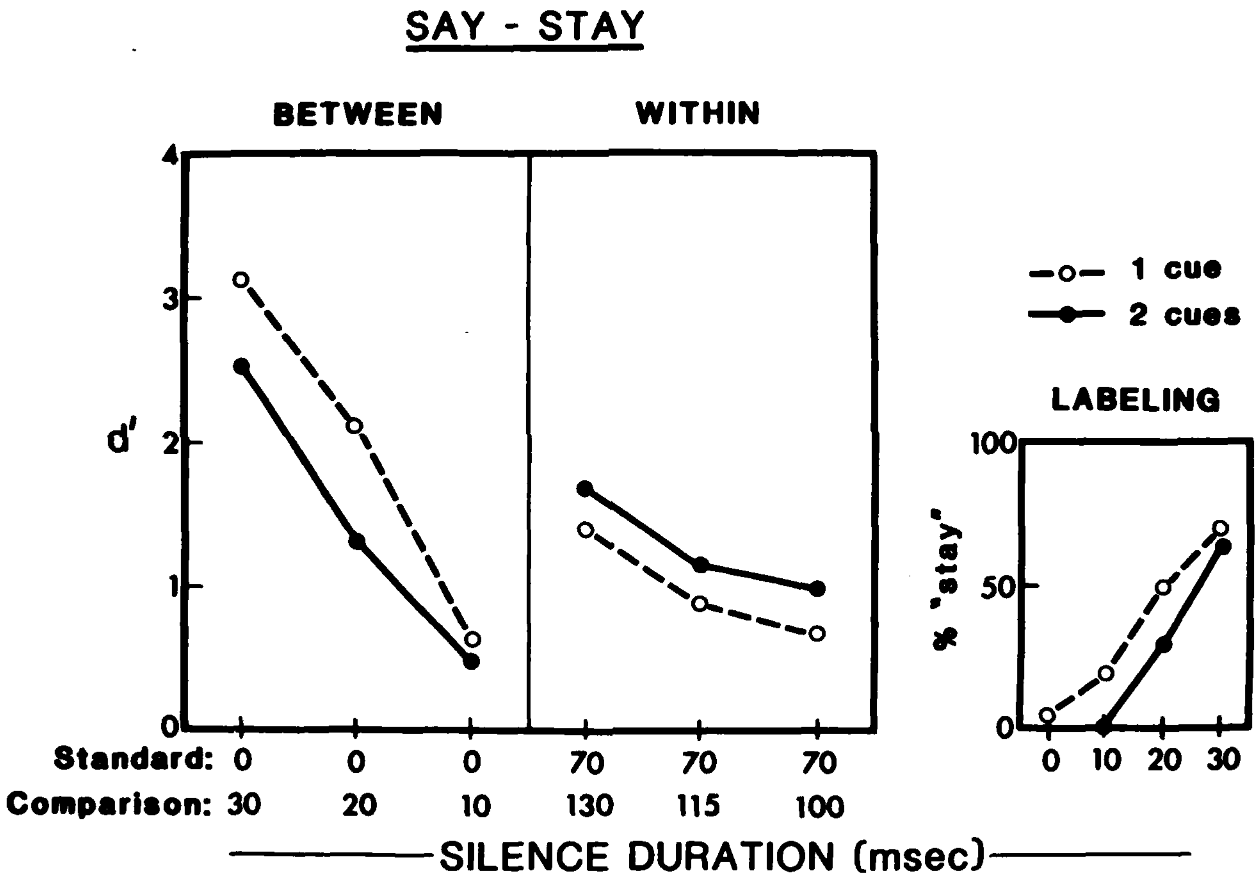


Figure 1. Results of Experiment 1.

Results

The results are shown in Figure 1. The first panel shows average d' scores in the Between condition. Discrimination performance was high in the first block but decreased rapidly as the difference in the primary cue was reduced, F(2,20) = 24.4, $p < .001$. As predicted from the trading relation between the primary and secondary cues, performance was higher on 1-cue than on 2-cue trials; however, this difference did not reach significance due to high intersubject variability, F(1,10) = 3.7, $p < .10$. The Blocks by Cues interaction was likewise nonsignificant.

The second panel of Figure 1 shows the results of the Within condition. These results represent the combined (i.e., averaged d') scores of the two repetitions of this condition, which exhibited highly similar response patterns. Performance was only slightly better in the second run, F(1,10) = 4.0, $p < .10$; no factor interacted with Repetitions. Discrimination scores started at a lower level in this condition than in the Between condition, even though the difference in the primary cue was twice as large. Performance declined over blocks, F(2,20) = 14.2, $p < .001$, and this effect did not interact with Cues. Most importantly, the difference between the two types of trials was reversed here, performance being better on 2-cue than on 1-cue trials, F(1,10) = 12.1, $p < .01$. This reversal was confirmed by a significant Conditions by Cues interaction in the joint analysis of the Between and Within conditions, F(1,10) = 6.6, $p < .05$.

The third panel of Figure 1 shows the labeling data for the stimuli used in the Between condition, obtained from a different group of subjects. One listener perceived all stimuli as "say" and was excluded. The data of the remaining eight listeners confirm that the standard stimulus (no silence) was heard as "say" and that the "say"-"stay" boundary fell between 20-25 msec, as expected. The labeling data also exhibit the trading relation between the two cues, with fewer "stay" responses to the 2-cue (i.e., conflicting-cues) stimuli. However, this difference once more did not reach significance because of high intersubject variability, F(1,7) = 4.0, $p < .10$.

Discussion

Basically, the results confirmed the predictions: A trading relation between the two cues appeared, though not very reliably, in the region of the "say"-"stay" boundary, whereas it was clearly absent within the "stay" category. This suggests, in accordance with the findings of Best et al. (1981), that the trading relation between silence duration and F1 onset frequency is phonetic, rather than auditory, in origin.

The present data are somewhat weakened by the nonsignificance of the trading relation in the Between condition and in the labeling task. However, we must also consider that (1) the difference in the secondary cue was rather small and (2) the stimuli were presented in a discrimination paradigm that may have facilitated the detection of auditory stimulus differences in the Between condition, even more so as this condition was preceded by the Within condition, which required auditory discrimination of similar differences. Any phonetic trading relation between the relevant cues (or, rather, its manifestation as superior performance on 1-cue trials) would be weakened by auditory discrimination beyond the detection of phonetic differences, since auditory discrimination bestows an advantage on 2-cue trials. Therefore, the critical result is the change across conditions in the relation between 1-cue and 2-cue discrimination--a change that was significant in the present experiment.

It is conceivable, of course, that an auditory trading relation between silence duration and F1 onset frequency exists when the silence is short but not when it is long. The most plausible form of this hypothesis would be that the presence of a silent interval is more difficult to detect when F1 has a higher onset, but that the perceived duration of longer silent intervals is not affected by F1 onset frequency. This hypothesis is consistent with the present data, but it seems unlikely in view of the Best et al. (1981) findings. Specifically, these authors found that subjects who perceived sine-wave analogs of "say"-"stay" stimuli nonphonetically and focused on the silence cue were not at all affected by F1(-analog) onset frequency, even when the silence durations were in the short range.

In the Best et al. study, it was found that listeners who followed an auditory strategy focused on one cue and ignored the other. In the present Within condition, selective attention to the silence cue would have resulted in equal scores on 1-cue and 2-cue trials, both declining over blocks, whereas selective attention to the spectral cue would have resulted in much better performance on 2-cue than on 1-cue trials, with no decline in 2-cue discrimination performance over blocks. However, no subject exhibited this second pattern, and few exhibited the first. Thus, the average data (Fig. 1) are fairly typical of the individual subject; they are not an artifact of averaging over subjects with radically different strategies. It seems likely, then, that the present subjects took both cues into account, even though the practice trials encouraged selective attention to the primary cue and subjects' reports indicated that they had little awareness of the (rather small) difference in the secondary cue. In that case, the higher scores on 2-cue than on 1-cue trials simply show that stimuli differing on two dimensions are easier to discriminate than stimuli differing on one dimension only, which is perfectly plausible and consistent with the relative auditory independence of the two cues shown by Best et al. (1981). Their finding that subjects paid selective attention to one or the other cue was probably due to their paradigm, an AXB classification task in which the two cues were perfectly correlated in the reference stimuli (A, B). Thus, their subjects were encouraged to select one cue and ignore the other, redundant one; in fact, this strategy simplified the subjects' task. The present AX discrimination task, on the other hand, while it emphasized the silence cue, encouraged listeners to pay attention to all possible stimulus differences. The ability of subjects to make use of both cue dimensions in one task is not inconsistent with their ability to select only one of them in a different task, since either strategy may be followed with independent auditory dimensions.

It should be noted that the advantage of 2-cue over 1-cue trials in the Within condition did not increase over blocks (as might be expected if subjects began to direct their attention to the secondary cue as the difference in the primary cue got smaller) but remained constant at about 0.3 d', which provides an estimate of the (rather poor) discriminability of the secondary-cue difference, assuming that the discriminabilities of the two cues were additive. Another feature of the data worth mentioning is the apparent convergence of the 1-cue and 2-cue scores in the last block of the Between condition. Although this effect was not significant, it was quite clearly exhibited by several individual subjects. Note that the phonetic trading relation between the cues is expected to disappear not only within the "stay" category but also within the "say" category—a situation approximated by the third block of the Between condition.

EXPERIMENT 2: "SAY SHOP"-"SAY CHOP"

The trading relation investigated in this experiment involved the same primary cue as in Experiment 1, viz., duration of silence, but a different secondary cue--the duration of the fricative noise following the silence. The trading relation between these two cues was demonstrated by Repp et al. (1978): More silence was needed to turn "say shop" into "say chop" when the fricative noise was long than when it was short.

This trading relation has much in common with that of Experiment 1; however, it does involve two cues varying along the same physical dimension (duration), which makes an auditory interaction perhaps more likely than between a temporal and a spectral dimension. For example, there may be a contrastive effect, such that a long fricative noise makes the preceding silence sound relatively short (or vice versa), which would lead to the observed trading relation. The present study put this hypothesis to test, using the same paradigm as Experiment 1. If there is an auditory interaction between the two temporal cues, then it should surface regardless of whether or not subjects perceive phonetic contrasts.

Method

Subjects. Ten volunteers participated, two of whom had also been subjects in Experiment 1, and six of whom had previously been subjects in Experiment 3b.

Stimuli. The stimuli were created on the OVE IIIc synthesizer. Formant parameters were copied from a spectrogram of "say shop" produced by a male speaker (as used in Repp et al., 1978). Synthetic stimuli were used because it turned out to be difficult to change the duration of a natural fricative noise without audible clicks or other discontinuities. The initial 240-msec "say" portion was followed by a variable silent interval, a fricative noise of variable duration, and a 125-msec final periodic portion ("op") whose first 10 msec overlapped the last 10 msec of the fricative noise. The fricative noise reached maximum amplitude after 50 msec. Fundamental frequency rose from 85 to 100 Hz during the "ay" portion and fell from 100 to 90 Hz during the "op" portion.

The primary cue was the amount of silence preceding the fricative noise. In the Between condition, the standard stimulus had no silence at all ("say shop"), and the comparison stimuli had 30, 20, and 10 msec, respectively, on "different" trials in the three test blocks, just as in Experiment 1. In the Within condition, the standard had 40 msec of silence ("say chop"), and the comparison values were 100, 80, and 60 msec. The "say shop"-"say chop" boundary was expected to be in the vicinity of 20 msec of silence. The

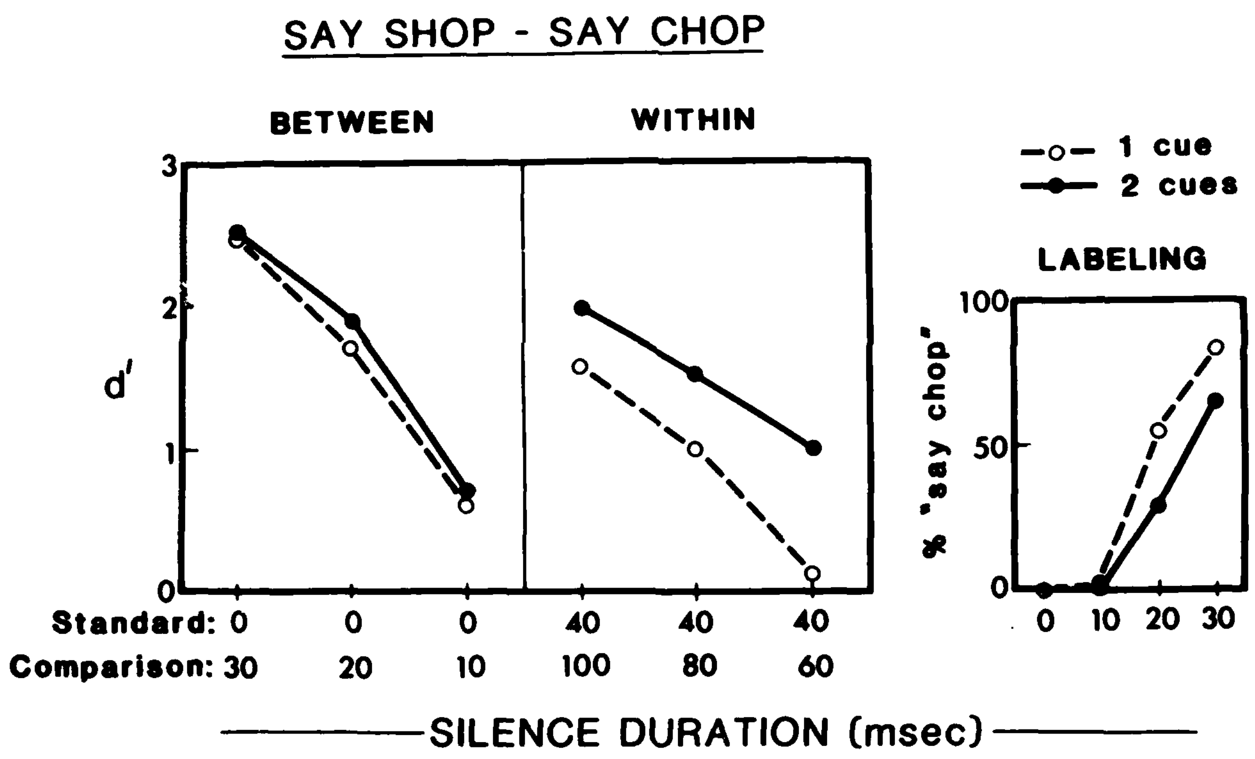


Figure 2. Results of Experiment 2.

secondary cue was the duration of the fricative noise in the second syllable. On 1-cue trials, its duration was 110 msec, whereas, on 2-cue trials, it was 130 msec, thus biasing perception more towards "say shop." The duration of the noise was changed at the synthesis stage by extending its central steady-state portion. The stimulus tapes were recorded directly from the synthesizer, without digitization of stimuli, so the fricative noise waveforms exhibited natural random variability across tokens.

Results

The results are shown in Figure 2. The first panel shows that the average performance level in the Between condition was similar to that in Experiment 1 (where the same values of silence had been employed), with a similarly striking decline over blocks, F(2,18) = 11.8, *p* < .001. However, there was no difference between 1-cue and 2-cue trials; in other words, the trading relation did not emerge.

In the Within condition (second panel of Fig. 2), performance was somewhat lower despite the larger differences in the primary cue. Performance declined over blocks, F(2,18) = 16.9, *p* < .001. In addition, however, accuracy on 2-cue trials was a good deal better than on 1-cue trials, F(1,9) = 32.3, *p* < .001. This difference seemed to increase over blocks, but the Cues by Blocks interaction did not reach significance. There was no significant effect involving Repetitions. The joint analysis of the Between and Within conditions revealed a significant Conditions by Cues interaction, F(1,9) = 22.4, *p* < .002, which confirmed the different effects that addition of a secondary cue had in the two conditions.

The labeling results (third panel of Fig. 2), obtained from the same group of subjects, revealed that the standard was always heard as "say shop" and that the phonetic category boundary fell between 20-25 msec, as expected. However, there was also the expected trading relation, with more "say chop" responses to stimuli containing the shorter noise, F(1,9) = 16.9, *p* < .01. Thus, the trading relation was exhibited in labeling but not in Between discrimination.

The reliability of the pattern of results shown in Figure 2 was confirmed by the results of the author and his research assistant who took the test as pilot subjects. Both showed the pattern in especially clear form: No trading relation in the Between condition but a large advantage for 2-cue trials in the Within condition.

Discussion

Except for the complete absence of a trading relation in the Between condition, the present data are quite similar to those of Experiment 1, suggesting that the trading relation between silence and fricative noise durations is similar to that between silence duration and F1 onset frequency, and that both are phonetic in origin. Both, of course, concern the perception of the same phonetic contrast--stop manner. As in Experiment 1, the critical finding is the Conditions by Cues interaction, which reflects the change in the difference between 1-cue and 2-cue trials across conditions. The absence of a trading relation in the Between condition is probably due to listeners' detection of auditory differences in addition to the phonetic contrast. Since the difference in the secondary cue was more noticeable here than in Experiment 1 (as suggested by the larger difference between 1-cue and 2-cue trials in the Within condition), the resulting auditory advantage for 2-cue trials may have completely canceled the advantage for 1-cue trials due to the phonetic trading relation in the Between condition.

The difference between the 1-cue and 2-cue d' functions in the Within condition suggests that the discriminability of the secondary cue difference was about 0.4 d' at the outset and increased to 0.9 d' in the last block, where discrimination on 1-cue trials was at chance. Although this increase did not reach significance, it does suggest that some subjects directed their attention towards the noise duration difference as the silence duration difference got smaller. The data also suggest, surprisingly, that the difference between a 110-msec and a 130-msec noise was much easier to detect than the difference between a 40-msec and a 60-msec silence (Within condition, last block). Since this finding contradicts Weber's Law, it indicates that silence and noise durations are not equivalently represented on the subjective temporal dimension.

An auditory hypothesis compatible with the present data would be that the detection of silence is not affected by the duration of a following noise segment, while the perceived duration of a longer silence is increased when the duration of the noise is increased. The direction of this hypothetical effect does not seem right, but at present there is no direct evidence against this hypothesis. The relevant psychoacoustic experiments remain to be done.

EXPERIMENT 3: "GOAT"-"COAT"

This study was concerned with a trading relation reported by Repp (1979): When voice onset time (VOT) is used as the primary cue to the voicing of an utterance-initial stop consonant, less increase in VOT is needed to turn a voiced stop into a voiceless one when the amplitude of the aspiration noise (whose duration is the VOT) is reduced. This trading relation is different in two important respects from those investigated in Experiments 1 and 2. First, the two interacting cues are both properties of the same signal portion, viz., of the aspiration noise that precedes voicing onset. Second, it appears that there is no good articulatory rationale for this trading relation. Although the relevant measurements have not been done, it seems likely that the amplitude of aspiration, measured at a fixed distance from the release, would be about the same in voiced and voiceless stops. It is true, of course, that voiced stops have a much shorter period of aspiration, and this necessary covariation of aspiration duration and time-integrated amplitude may be sufficient to account for the perceptual trading relation. Still, the articulatory explanation seems less compelling than that for other effects, where different cues can be shown to be acoustically diverse consequences of the same articulatory act (cf. Repp et al., 1978). Moreover, there are well-known instances of trade-offs between duration and amplitude at the auditory threshold and in judgments of loudness (e.g., Garner & Miller, 1947; Small, Brandt, & Cox, 1962). For these reasons, the present trading relation may well be auditory in origin. If so, it was predicted to occur in both conditions of Experiment 3; that is, performance was expected to be higher on 1-cue than on 2-cue trials in both the Between and Within conditions.

Experiment 3 was run twice. The first run (Exp. 3a) was only partially successful because the stimuli in the Between condition turned out to have missed the boundary (their VOTs were too long), so that the Between condition was effectively another Within condition. Also, the VOT differences were rather small, so that the subjects were in great trouble. Therefore, a replication (Exp. 3b) was conducted with shorter VOT values in the Between condition and larger VOT differences. Results from both runs will be reported. The labeling test was administered at the end of Experiment 3b.

Method

Subjects. Eight volunteers participated in Experiment 3a. All of them had previously been subjects in Experiment 1. There were nine subjects in Experiment 3b, two of whom had also been in Experiment 3a.

Stimuli. In contrast to the previous stimuli, the present ones were modified natural speech. A female speaker recorded the words "goat" and "coat." They were digitized at 10 kHz, and a VOT continuum was constructed by first replacing the burst and aspiration portions of "goat" (22 msec) with the first 22 msec of "coat" and by then substituting additional equivalent amounts of aspiration noise from "coat" (VOT = 66 msec) for each successive pitch period of "goat." For a detailed description of this procedure, see the appendix in Ganong (1980).

Stimuli from this continuum were used in the Between condition only. For the Within condition, where VOTs longer than that of the natural "coat" were required, the stimuli were generated by a different procedure. Note that, in the method described above, total stimulus duration remains constant as VOT is increased while the periodic stimulus portion is progressively shortened. This is standard procedure for VOT continua and probably does not matter when relatively short VOTs are to be discriminated. However, when VOTs are made rather long, little is left of the periodic portion, and informal observations have shown that removal of even a single pitch period may become perceptually quite salient. That is, subjects may discriminate such stimuli not on the basis of VOT but on the basis of changes in the duration and intonation of the "vowel." To prevent this from happening in the present Within condition, the periodic stimulus portion was held constant, and VOT was further increased by duplicating randomly selected segments of the final portion of the aspiration noise, where the formant transitions presumably were close to asymptote.

Thus, the stimuli in the Between condition had a total duration of 228 msec (VOT plus periodic portion), with the periodic portion diminishing as VOT increased, whereas the stimuli in the Within condition had a constant periodic portion of 155 msec, and total duration increased with VOT. All stimuli included, in addition, a rather powerful final [t] release burst of approximately 112 msec duration, which was separated from the end of the periodic portion by a 133-msec silent closure interval.

The primary cue in this study was, of course, VOT (i.e., the duration of the aperiodic portion at stimulus onset). In the Between condition of Experiment 3b (that of Experiment 3a will not concern us here, since performance was at chance), the standard had a VOT of 38 msec (which seems rather long but was still heard as "goat"), and the comparison stimuli had

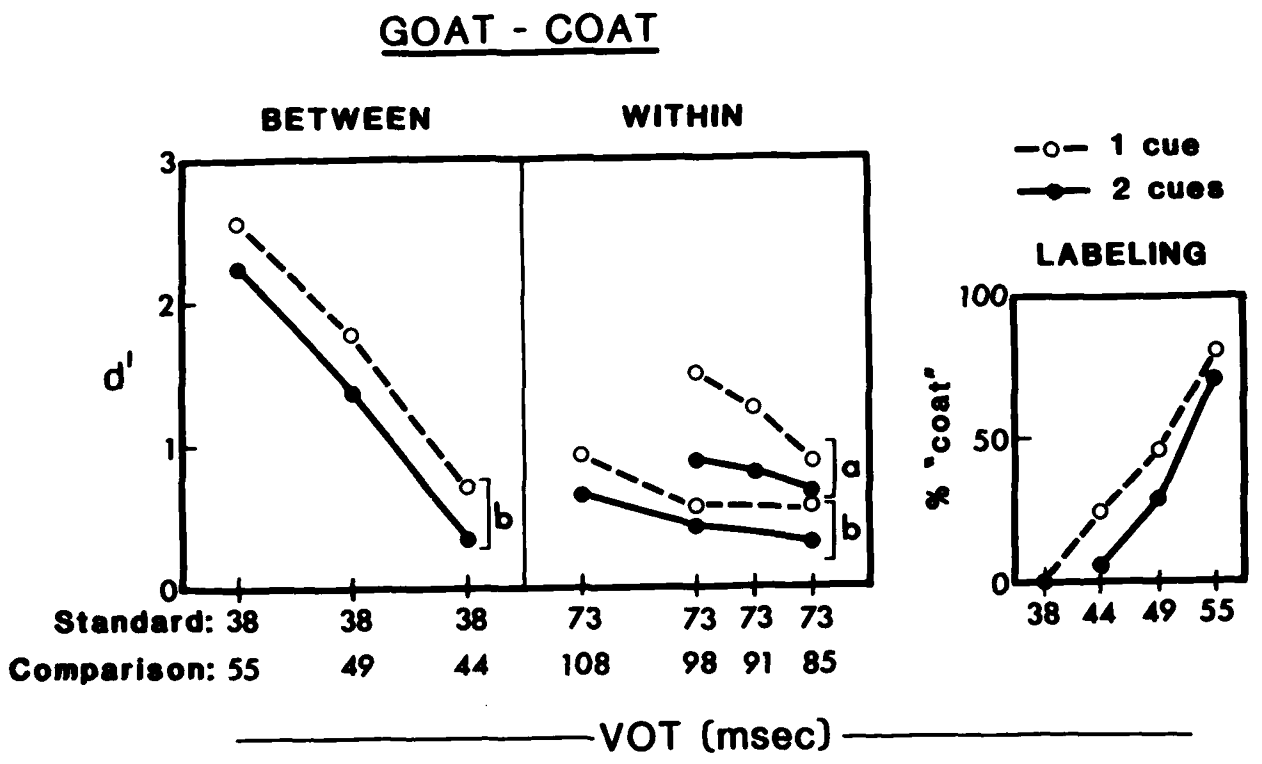


Figure 3. Results of Experiments 3a and 3b.

VOTs of 55, 49, and 44 msec, respectively. In the Within condition of Experiment 3b, the standard had a VOT of 73 msec ("coat"), and the comparison stimuli had values of 108, 98, and 85 msec, respectively. In Experiment 3a, the same standard was used, but the comparison stimuli had values of 98, 91, and 85 msec. The secondary cue was the amplitude of the aperiodic stimulus portion. On 2-cue trials, it was reduced by 6 dB SPL in the comparison stimulus, counteracting the longer VOT of that stimulus. This manipulation was performed on the digitized waveform, using computer instructions.

Results

Within-category discrimination of the "goat"-"coat" stimuli proved to be a difficult task for naive subjects. One problem seemed to be to discover the dimension on which the stimuli differed. (Recall that the nature of the difference was not revealed in the instructions but had to be detected during the practice block.) In Experiment 3a, performance in the first presentation of the Within condition was close to chance (average d' = 0.31), and there was no difference between 1-cue and 2-cue trials. A similar result was obtained in the Between condition where, because of inappropriately long VOTs, all but one subject heard only "coat" and performed at chance level. The single subject who appeared to be able to make use of phonetic contrasts performed quite well and had higher scores on 1-cue than on 2-cue trials, in accord with the expected trading relation. Prompted by subjects' complaints over the difficulty of the task, the experimenter told them before the repetition of the Within condition what kind of difference to listen for, and he produced exaggerated examples of stops with different amounts of aspiration to illustrate the point. This had a striking effect on (most) subjects' performance. The results from this final condition of Experiment 3a are presented in the second panel of Figure 3 (the functions labeled "a"). It can be seen that performance was better on 1-cue than on 2-cue trials, F(1,7) = 5.7, $p < .05$. This pattern contrasts with that obtained in the Within conditions of Experiments 1 and 2, where the opposite difference was observed. Due to large variability, neither the decline in performance across blocks nor the Blocks by Cues interaction reached significance.

The subjects in Experiment 3b were told right at the outset to direct their attention to the initial portion of the stimuli; however, they were not told the precise nature of the difference to listen for. Surprisingly, the hint did not help. Performance in the first Within condition was poor, despite the increased VOT differences (average d' = 0.23), and there was no clear difference between 1-cue and 2-cue trials. Therefore, these data were again discarded. However, the choice of VOT values for the Between condition was more successful this time; these results are shown in the first panel of Figure 3. Subjects performed at a level comparable to that in Experiments 1 and 2, although the durational differences were somewhat smaller here. Performance declined over blocks, F(2,16) = 5.6, $p < .05$. Scores were higher on 1-cue than on 2-cue trials, F(1,8) = 5.5, $p < .01$, which reflects the expected trading relation.

The results of the repetition of the Within condition are shown in the second panel of Figure 3 (labeled "b"). These subjects, too, were told what difference to listen for before they repeated the Within condition. However, their performance improved less than that of the subjects in Experiment 3a.

Although better than chance, on the average, scores were low and highly variable. Neither the Blocks effect nor the Cues effect was significant; note, however, a tendency for 1-cue discrimination to be higher than 2-cue discrimination. This tendency is supported not only by the results of Experiment 3a but also by the data of a research assistant who served as a pilot subject and showed a striking advantage for 1-cue trials in both conditions. The Cues by Conditions interaction was nonsignificant.

The third panel of Figure 3 shows labeling data deriving from six of the subjects plus the research assistant. (Three subjects had already been tested before it was decided to add the labeling test.) These data confirm that the standard stimulus (VOT = 38 msec) was perceived as "goat," and they also show the expected trading relation, although it fell short of significance, $F(1,6) = 5.5$ $\underline{p} < .10$.

Discussion

The results of this experiment are stronger in terms of what they do not show than in what they do show. The most significant finding is the absence of an advantage for 2-cue trials in the Within condition. The data suggest that, on the contrary, there was an advantage for 1-cue trials in both the Within and Between conditions. This pattern of results is the one expected for a trading relation of psychoacoustic origin. The interaction between aspiration noise duration and amplitude may be similar to other kinds of auditory time-intensity trade-offs.

EXPERIMENT 4: "CHOP"-"SHOP"

The trading relation studied in this last experiment has been known for a long time: It concerns fricative noise duration and rise-time (i.e., the time from noise onset to the point of maximum amplitude) as joint cues to the fricative-affricate distinction. Gerstman (1957) showed that, to turn an utterance-initial [ʃ] into a [tʃ], the noise duration needs to be shortened more if its rise-time is slow; or, conversely, its rise-time must be shortened more if noise duration is long. Gerstman excluded the rise-time portion from his measure of noise duration, thus confounding total noise duration with the rise-time variable. Van Heuven (1979) recently reanalyzed Gerstman's data and found that total noise duration accounted for nearly all the variance; rise-time made only a small contribution to perception. Still, it can hardly be doubted that amplitude rise-time has some cue value for the fricative-affricate distinction. Although some relevant studies have confounded rise-time with amplitude at onset, which itself may be an important cue (e.g., Dorman, Raphael, & Liberman, 1979: Exp. 5), others have shown rise-time proper to be a sufficient cue (e.g., Cutting & Rosner, 1974; Rosen & Howell, 1981). Thus, it seems likely that rise-time can be traded against noise duration, at least within certain limits.

Like the trading relation investigated in Experiment 3, that between the present two cues engages two properties of the same signal portion. It is possible that these properties interact at the auditory level to determine the perceived duration of the noise, or possibly its perceived abruptness of onset. However, the present trading relation, unlike that of Experiment 3, also has a good articulatory explanation: Naturally produced fricatives and affricates differ in both noise duration and rise-time. Experiment 4 was expected to shed light on the origin of this trading relation.

Experiment 4 actually consisted of two experiments, identical except for the stimuli. In Experiment 4a, the full "chop"-"shop" stimuli were used. In Experiment 4b, only the fricative noise portions were presented. This second experiment was intended to serve as a kind of nonspeech control for the first, since informal observations had suggested that the isolated fricative noises did not invite phonetic categorization as "sh" or "ch," or in any case were more difficult to label than the full stimuli. It was expected that whatever phonetic effects might be present in the Between condition of Experiment 4a would be absent in the corresponding condition of Experiment 4b.

Method

Subjects. Nine volunteers participated, five of whom had also been subjects in earlier experiments. All subjects took Experiment 4a first, then Experiment 4b on a separate day.

Stimuli. The stimuli were created on the OVE IIIc synthesizer; they were derived from the second halves of the stimuli of Experiment 2. The choice of cue values for the Between condition was guided by Gerstman's (1957) data. The primary cue was fricative noise duration. In the Between condition, the duration was 70 msec for the standard (intended to be heard as "chop") and 100, 90, and 80 msec, respectively, for the comparison stimuli. In the Within condition, the standard had a 140-msec noise ("shop"), and the comparison values were 200, 180, and 160 msec. The secondary cue was the rise-time of the noise. On 1-cue trials, it was 50 msec; on 2-cue trials, it was reduced to 30 msec (favoring "chop" percepts). In each case, the amplitude rise was linear and onset amplitude was set at the minimum value possible in synthesis; amplitude parameter values for the two different rise-times began to diverge after the initial 5 msec. The accuracy of the rise-times was verified by digitizing and displaying the waveforms of the stimuli. Stimulus tapes were recorded directly from the synthesizer to avoid artifacts due to "frozen" noise waveforms.

Results

The results of Experiment 4a are the functions labeled "a" in Figure 4. Performance in the Between condition was again comparable to that in previous experiments; the decline over blocks was significant, $F(2,16) = 13.0$, $p < .001$. However, there was no difference between 1-cue and 2-cue trials. A slight avantage for 1-cue trials at the outset changed to a slight advantage for 2-cue trials in the last block, but the Blocks by Cues interaction was not significant.

Surprisingly, the results of the Within condition were remarkably similar to those of the Between condition. There was no significant effect involving the Repetitions factor. Performance declined over blocks, $F(2,16) = 26.5$, $p < .001$, and an advantage for 2-cue trials emerged in the second and third blocks. The Blocks by Cues interaction reached significance here, $F(2,16) = 4.2$, $p < .05$. This interaction was also obtained in the joint analysis of the Between and Within conditions, $F(2,16) = 7.4$, $p < .01$, with no triple

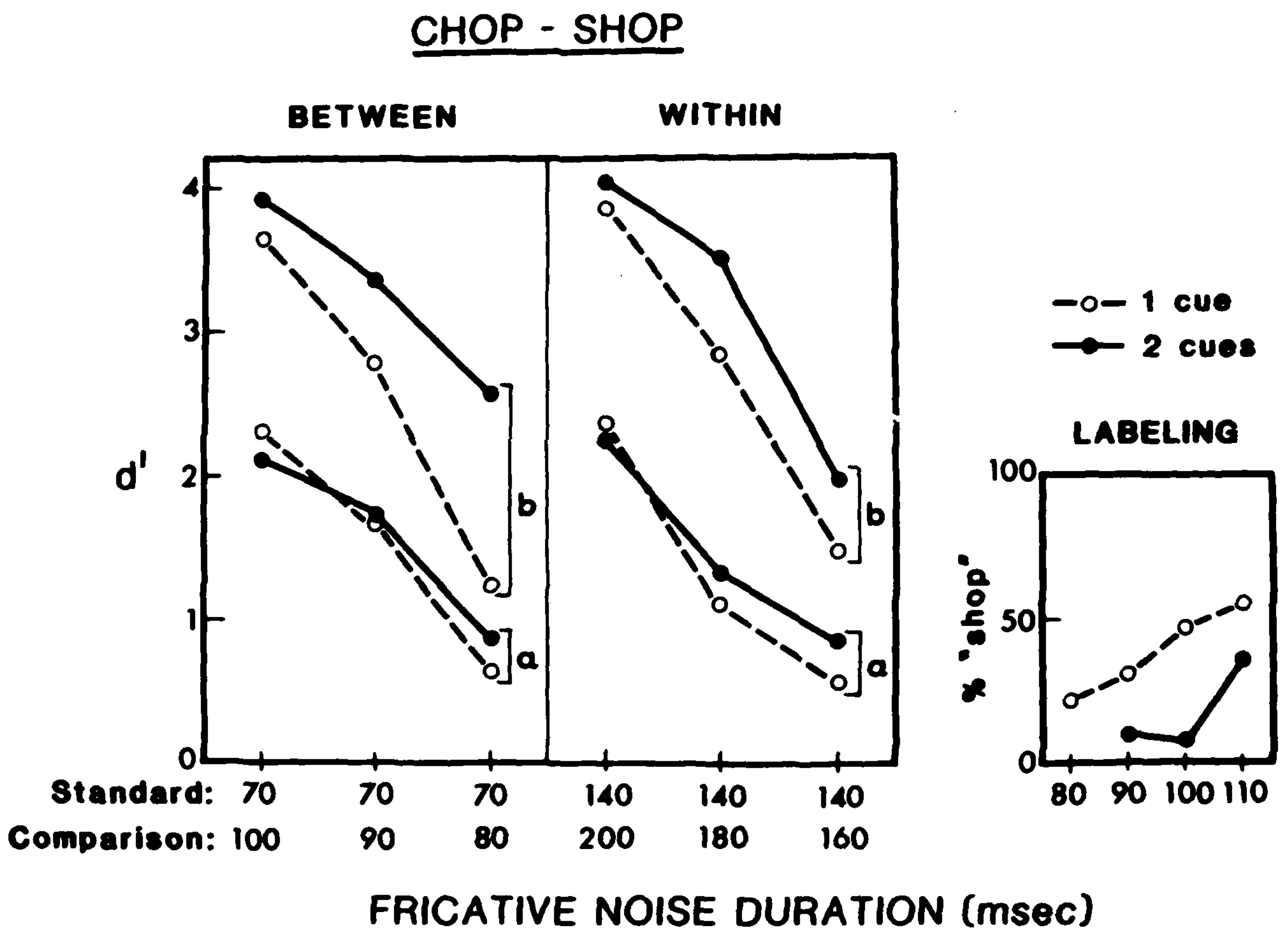


Figure 4. Results of Experiments 4a and 4b.

interaction involving Conditions, which confirms the similarity of the response patterns in the two conditions.

The labeling data were less tidy than in the earlier experiments; in particular, the standard stimulus was not an unequivocal "chop" for all listeners. However, the trading relation between the noise duration and rise-time cues was present and significant, $F(1,8) = 21.0$, $p < .01$.

The results of Experiment 4b (fricative noise portions only) are labeled "b" in Figure 4. Performance was strikingly better here than in Experiment 4a. Also, in contrast to Experiment 4a, a large advantage for 2-cue trials can be seen, both in the Between condition, $F(1,8) = 19.7$, $p < .01$, and in the Within condition, $F(1,8) = 47.9$, $p < .001$. The results had in common with those of Experiment 4a the Blocks by Cues interaction: The advantage for 2-cue trials increased over blocks, particularly in the Between condition, $F(2,16) = 11.3$, $p < .001$. The interaction did not reach significance in the Within condition, where it may have been due to a ceiling effect in Block 1. The different patterning of this interaction in the two conditions was reflected in a significant Conditions by Blocks by Cues interaction, $F(2,16) = 6.4$, $p < .01$. There was no effect involving Repetitions in the Within condition.

Discussion

The "chop"-"shop" stimuli were the most problematic ones of the present set. Not only was the phonetic contrast less clear-cut, but the author also noted as a pilot subject that the stimuli were prone to auditory segregation: After some minutes of listening, the fricative noise would suddenly "stream away" from the periodic portion, thereby destroying the speechlikeness and perceptual coherence of the stimuli. These observations are in accord with the results, which show little difference between the Between and Within conditions, suggesting that listeners may have made little or no use of phonetic labels in Between discrimination. The Blocks by Cues interaction may indicate that subjects made some use of phonetic labels in the first block of both conditions and abandoned this strategy later. This is not implausible in view of the possibility that the standard stimulus in the Within condition may not have been an unequivocal "shop"; it is also supported by the reports of some subjects who claimed to have heard a [ʃ]-[tʃ] contrast in the Within condition. However, this interpretation is called into question by the existence of a similar Blocks by Cues interaction in Experiment 4b, where phonetic labeling presumably played no role. We may presume, then, that the interaction reflects a change in auditory strategies: As long as differences in noise duration were large, listeners paid attention to that cue dimension, and only as the differences got smaller was their attention directed to the rise-time differences as well.

Two aspects of the present results are clear. First, fricative noise duration and rise-time do not seem to engage in an auditory trading relation; otherwise, an advantage for 1-cue trials should have been observed in the Within condition, just as in Experiment 3. Therefore, the trading relation observed in the labeling task is likely to be phonetic in nature, and its failure to show up in Between discrimination may be ascribed to procedural factors and to the above-mentioned stimulus problems. Second, the periodic portion of the "chop"-"shop" stimuli seemed to interfere with auditory memory for the duration of the fricative noise, or with the perception of that duration in the first place: Discrimination was considerably easier when the noises were presented in isolation. Perhaps, this difference reflects differently-sized auditory units; it might disappear when the noise is perceptually segregated from the periodic portion, either as the consequence of prolonged listening or of a listener-controlled strategy. However, Repp (1981a) found that isolated fricative noises differing in spectrum (rather than duration) were more accurately discriminated in isolation than when followed by a periodic portion, even by subjects who were able to perceptually segregate the noise from the periodic portion. Thus, even though the stimulus components could be isolated by perceptual strategies, they were not completely independent in auditory memory.

GENERAL DISCUSSION

Even though the present results must be considered preliminary, they are encouraging, and the technique used promises to provide a relative effortless way of determining the origin of a trading relation. The postexperimental labeling tests showed the expected trading relations in all cases (although it was not statistically reliable in two). Thus, the stimuli seemed appropriate, even though they had not been formally pretested. However, the expected trading relations were not consistently present in the Between discrimination conditions. In two studies ("say"-"stay," "goat"-"coat"), they showed up, but not very reliably; in the other two ("say shop"-"say chop," "chop"-"shop"), they were definitely absent. The proposed reason for this was that the fixed-standard AX paradigm encouraged listeners to make maximal use of whatever auditory differences they could detect between the stimuli. For example, Carney, Widin, and Viemeister (1977) and Ganong (1977) successfully used the same paradigm to get subjects to discriminate small differences in VOT within a phonetic category. Auditory discrimination, in addition to discrimination based on phonetic labels, would tend to reduce the trading relation observed in the Between condition, unless the trading relation itself is of auditory origin. It also seems that differences in fricative noise duration were relatively salient, which may explain the absence of an advantage for 1-cue trials in the Between conditions for both "say shop"-"say chop" and "chop"-"shop."

The critical data came from the Within conditions of the different experiments. In two studies ("say"-"stay," "say shop"-"say chop"), there was an advantage for 2-cue trials, which contrasted with the pattern of results in the Between condition. This outcome suggests strongly that the trading relations between the relevant cues are phonetic in origin, confirming earlier results by Best et al. (1981) for the "say"-"stay" contrast. These trading relations--between silent closure duration and F1 onset frequency in the case of "say"-"stay," and between silent closure duration and fricative noise duration in the case of "say shop"-"say chop"--are well explained by reference to articulation, since in each case changes in the two cues are tightly correlated in the production of the relevant phonetic contrast. In a third study ("chop"-"shop"), the results were more ambiguous because similar results were obtained in the Between and Within conditions, and the advantage for 2-cue trials was not as clear-cut. However, since a clear trading relation was obtained in the labeling task, the trading relation is likely to be of phonetic origin. The articulatory rationale applies here, too: Both fricative noise duration and rise-time change together in the production of the fricative-affricate contrast. Thus, three of the trading relations investigated appear to be phonetic in nature, and each of them has an articulatory explanation.

Only the "goat"-"coat" stimuli yielded a different pattern. Here, there was an advantage for 1-cue trials in both the Between and Within conditions, suggesting an auditory origin for this trading relation. Significantly, this trading relation is also the only one that has no obvious articulatory correlates: Aspiration amplitude per se does not seem to vary in the voicing contrast for stop consonants. Thus, the present results fit the predicted pattern: A trading relation is phonetic in origin if it has articulatory correlates, but auditory in origin if it does not.

The results of the Within conditions also tell us something about the auditory perception of speech parameters. In some cases ("say"-"stay," "say shop"-"say chop," isolated noises of "chop"-"shop"), the two cue dimensions seemed to be independent and simultaneously accessible to the subjects. In the case of "goat"-"coat," on the other hand, they seemed to interact. This difference is reminiscent of the distinction between "separable" and "integral" stimulus dimensions (Garner, 1974; Lockhead, 1979). Integral dimensions are those where, in order for one dimension to exist, the other must be specified, and where selective attention to one dimension alone is not possible (Garner, 1974). Aspiration noise duration and amplitude seem to fit that description. However, the pairs of cues involved in the "say"-"stay" and "say shop"-"say chop" distinctions do not; they seem to be be separable at the auditory level, perhaps because they are also separated in time. In order to prove their auditory separability, it would be necessary to show that they can be selectively attended to, as Best et al. (1981) have done for "say"-"stay." The present task did not require selective attention, although it permitted such a strategy; the subjects, however, seemed to pay attention to both cue dimensions, which is certainly an option with separable cues. It is not clear where the "chop"-"shop" results stand in that regard; they are the ones most in need of replication.

Even though two cues may be auditorily separable, it is significant that they can nevertheless be integrated into a single phonetic percept. Presumably, this is achieved by a higher-level, speech-specific process that combines cues according to implicit knowledge about the articulatory and/or acoustic patterns of speech. It is not necessary to envision this process as one of cue extraction followed by cue recombination according to certain rules (the traditional machine metaphor); more vaguely, but probably more appropriately, it may be understood as a consequence of perceiving articulatory change through the acoustic signal, and of referring the perceived changes to internal criteria that specify the phonetic categories of the language. If so, it seems likely that attempts to explain phonetic trading relations by auditory psychophysics will, in most cases, remain futile.

REFERENCE NOTES

1. Miller, J. L., & Eimas, P. D. *Biological constraints on the acquisition of language: Further evidence from the categorization of speech by infants.* Manuscript submitted for publication, 1981.

2. Espinoza-Varas, B. *Optional processing strategies in multidimensional discrimination: Integration versus selective attention.* Paper presented at the 97th Meeting of the Acoustical Society of America, Cambridge, MA, June 1979.

REFERENCES

Bailey, P. J., Summerfield, Q., & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report on Speech Research*, 1977, *SR-51/52*, 1-25.

Best, C. T., Morrongiello, B., & Robson, R. The perceptual equivalence of two acoustic cues for a speech contrast is specific to phonetic perception. *Perception & Psychophysics*, 1981, *29*, 191-211.

Carney, A. E., Widin, G. P., & Viemeister, N. F. Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, 1977, *62*, 961-970.

Creelman, C. D. Human discrimination of auditory duration. *Journal of the Acoustical Society of America*, 1962, *34*, 582-593.

Cutting, J. E., & Rosner, B. S. Categories and boundaries in speech and music. *Perception & Psychophysics*, 1974, *16*, 564-570.

Dorman, M. F., Raphael, L. J., & Liberman, A. M. Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, 1979, *65*, 1518-1532.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, 1980, *27*, 343-350.

Ganong, W. F. III. *Selective adaptation and speech perception*. Unpublished doctoral dissertation, M.I.T., 1977.

Ganong, W. F. III. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, *6*, 110-125.

Garner, W. R. *The processing of information and structure*. Potomac, MD: Erlbaum, 1974.

Garner, W. R., & Miller, G. A. The masked threshold of pure tones as a function of duration. *Journal of Experimental Psychology*, 1947, *37*, 293-303.

Gerstman, L. *Cues for distinguishing among fricatives, affricates, and stop consonants*. Unpublished doctoral dissertation, New York University, 1957.

Heuven, V. J. van The relative contribution of rise time, steady time, and overall duration of noise bursts to the affricate-fricative distinction in English: A re-analysis of old data. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers presented at the 97th Meeting of the Acoustical Society of America*. New York: Acoustical Society of America, 1979.

Kuhl, P. K., & Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the*

Acoustical Society of America, 1978, 63, 905-917.

Liberman, A. M., & Studdert-Kennedy, M. Phonetic perception. In R. Held, H. Leibowitz, & H.-L. Teuber (Eds.), Handbook of sensory physiology, Vol. VIII. Heidelberg: Springer-Verlag, 1978.

Lockhead, G. R. Identification and the form of multidimensional discrimination space. Journal of Experimental Psychology, 1970, 85, 1-10.

Pastore, R. E. Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), Perspectives on the study of speech. Hillsdale, N. J.: Erlbaum, 1981.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.

Repp, B. H. Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants. Language and Speech, 1979, 22, 173-189.

Repp, B. H. Two strategies in fricative discrimination. Perception & Psychophysics, 1981, 30, 217-227. (a)

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Haskins Laboratories Status Report on Speech Perception, 1981, SR-67/68, this volume.(b)

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.

Rosen, S., & Howell, P. Plucks and bows are not categorically perceived. Perception & Psychophysics, 1981, 30, 156-168.

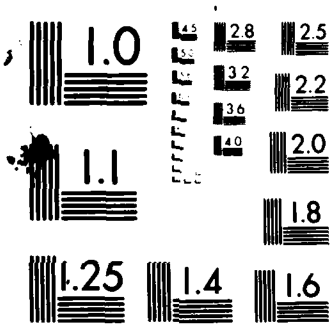
Small, A. M., Jr., Brandt, J. F., & Cox, P. G. Loudness as a function of signal duration. Journal of the Acoustical Society of America, 1962, 34, 513-514.

Summerfield, A. Q., & Haggard, M. P. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 1977, 62, 435-448.

AD-A111 385 HASKINS LABS INC NEW HAVEN CT F/G 6/16
STATUS REPORT ON SPEECH RESEARCH, JULY-DECEMBER 1981.(U)
1981 PHS-HD-01994
UNCLASSIFIED SR-67/68(1981) NL

3 · 3

END
DATE
FILMED
4 82
DTIC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS 1963 A

PRODUCTION AND PERCEPTION OF PHONETIC CONTRAST DURING PHONETIC CHANGE*

Paul J. Costa+ and Ignatius G. Mattingly

Abstract. Ten productions of each of the two words cod and card in the southeastern subdialect of the Eastern New England dialect, said to differ phonetically only in vowel length (and a number of foil words involving other phonetic contrasts) were recorded in a neutral carrier sentence by each of nine phonetically naive urban-Eastern New England speakers unaware of the purpose of the investigation. Spectrographic measurements revealed fairly consistent differences in the vocalic segment durations of cod and card for most speakers. But no speaker could reliably identify his own intended productions (though identification of foils was perfect). Evidently a phonetic change is in progress, and our results suggest that during such a change, contrasts in production may persist after they have ceased to be perceptually relevant.

It is usually taken for granted in phonetics that given a regular alternation in the production of two distinct lexical items, these two items will be perceived as different. Labov, Yaeger, and Steiner (1972), however, have reported instances of "partial mergers," in which, despite consistent acoustic differences between two phonetic types in a dialect, speakers of the dialect failed informal commutation tests. The purpose of this study is to examine another such situation in which the common assumption seems to be contradicted. The case in point is taken from the southeastern subdialect of the Eastern New England dialect—henceforth SENE—spoken in and around Fall River, Massachusetts.

It has been stated by Thomas (1958) and by Kenyon (1937), on the basis of data collected in the 30's for the Linguistic Atlas of New England, that in the case of low vowels, vowel length is distinctive for SENE. For example, there is said to be only a vowel length distinction between the two words cod [kɑd] and card [kɑ:d]. This implies that the acoustic signals for such pairs may differ solely in the duration of the vocalic segment. We have found, however, that while there is a fairly reliable durational difference in the production of cod and card, speakers of this dialect cannot consistently label their own productions. In other words, the distinction in production is virtually ignored in perception.

*Paper presented at 101st meeting of the Acoustical Society of America, at Ottawa, Canada, May 21, 1981.

+Department of Linguistics, University of Connecticut.

Acknowledgment. Support for the preparation of this paper was provided by Grant PFR-8006144 from the National Science Foundation.

[HASKINS LABORATORIES: Status Report on Speech Research SR-67/68 (1981)]

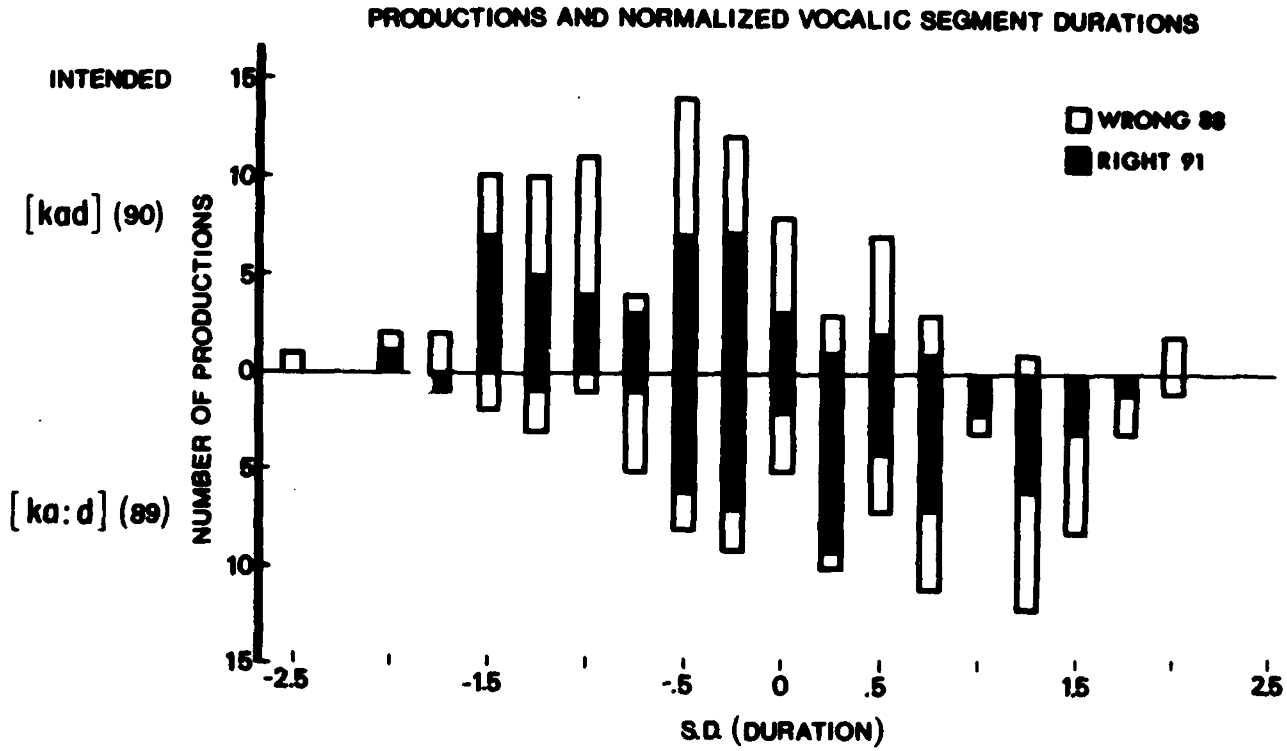

PRODUCTIONS AND NORMALIZED VOCALIC SEGMENT DURATIONS
INTENDED
[kad] (90)
[ka:d] (89)
NUMBER OF PRODUCTIONS
15
10
5
0
5
10
15
□ WRONG 88
■ RIGHT 91
-2.5
-1.5
-.5
0
.5
1.5
2.5
S.D. (DURATION)

Figure 1

We recently carried out a production experiment and a perception experiment with nine SENE speakers. In the production experiment, we collected acoustic data from the informants with which the acoustic correlates of the vowel length distinction between cod and card could be measured. The materials for each production experiment consisted of five sets of two minimally paired common English words differing with respect to a single feature: goat-coat, grate-grade, spit-spin, bit-pit, and cod-card. The first four pairs acted as foils. Each word was put into one of three carrier sentences so that it would be spoken at a natural tempo. A list in which all sentences appeared ten times, in random order, was prepared. The subject's task was to speak each sentence at a normal tempo. These utterances were recorded.

In a subsequent meeting with each subject, which took place from one hour to two days after the production experiment, a perception test was given. The stimuli for each subject were the 100 sentences he had spoken in the production experiment. Word pairs other than card and cod remained as foils. Each subject was asked to write down the test word in each sentence.

Wide band spectrograms were made of the ten tokens of cod and the ten tokens of card as spoken by each subject. Three successive durational measurements were noted: 1) the voice onset time for [k], 2) the vocalic duration measured from voice onset to the [d] closure, and 3) the closure duration for [d]. For each speaker the VOT and closure duration varied from token to token without a consistent pattern. On the other hand, the time measurements for the vocalic duration formed a rather consistent pattern. Measurements of vocalic duration plus VOT, or closure duration, or both, were less consistent than vocalic duration alone. Vocalic duration averages for the ten speakers for cod ranged from 210 to 320 msec, while averages for card ranged from 240 to 400 msec. The difference in the speaker average ranged from 30 to 40 msec. In all, three subjects made a definite split in their productions, four subjects were moderately consistent, and two were very inconsistent.

In order to pool the data in a way that would exclude, so far as possible, intersubject variation in speaking rate, we represented the vocalic duration of each token in signed units of standard deviation, using the average of each subject's durations for both cod and card as the mean for that subject. Thus, if each subject had produced all his tokens of card with longer durations than any of his tokens of cod, all card tokens would have greater signed values of standard deviation than any cod token.

Figure 1 shows the data pooled in this way. The number of cod productions for a particular range of standard deviation values is plotted as a histogram above the horizontal axis. In the same way, card productions are plotted below the horizontal axis. While there is a substantial overlap, it is clear that the proportion of cod productions decreases, and the proportion of card productions increases, as the standard deviation goes from extreme negative values (corresponding to relatively short durations) at the left, to extreme positive values (corresponding to relatively long durations) at the right. Thus the production data are consistent with the vowel length distinction described by Thomas and by Kenyon: the two words do differ in vocalic duration in production.

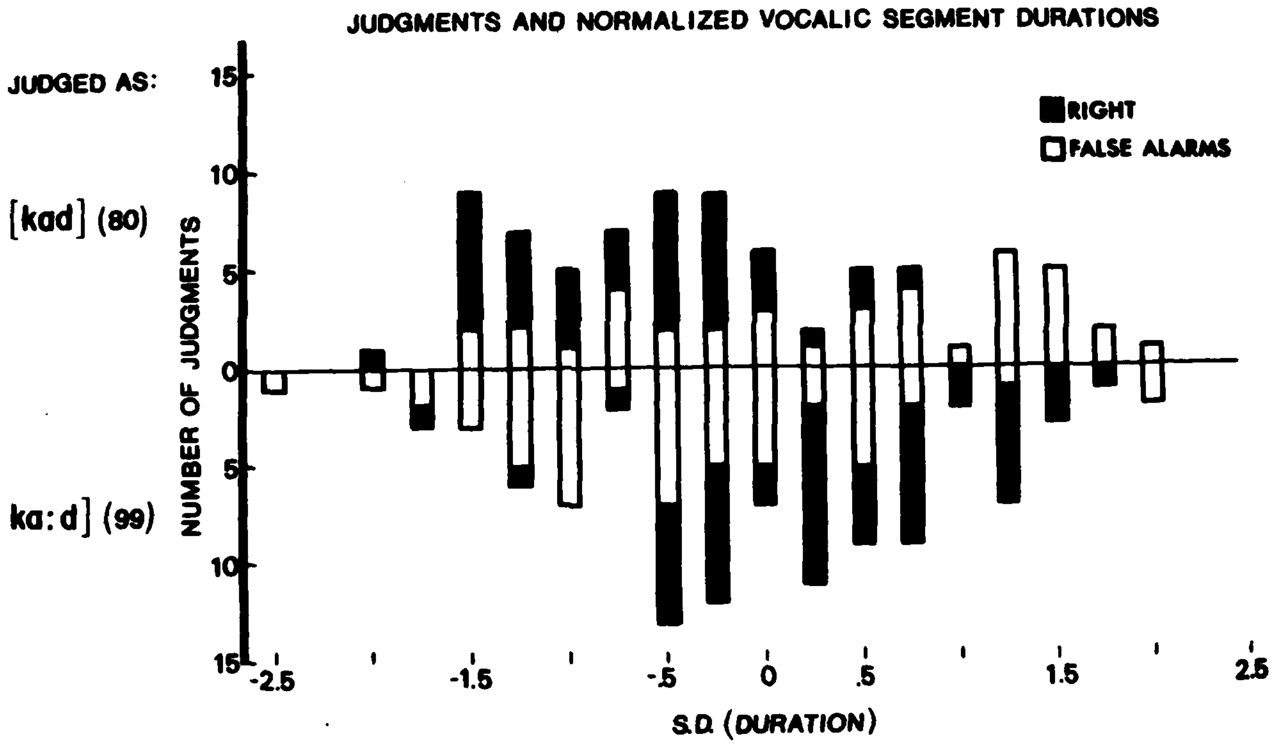

JUDGMENTS AND NORMALIZED VOCALIC SEGMENT DURATIONS
JUDGED AS:
[kad] (80)
ka:d] (99)
NUMBER OF JUDGMENTS
15
10
5
0
5
10
15
■RIGHT
□FALSE ALARMS
-2.5
-1.5
-.5
0
.5
1.5
2.5
S.D. (DURATION)

Figure 2

Perception is a different matter. Individual labeling results break the subjects down into three groups: two speakers with relatively consistent perception; four with inconsistent perception, and three with an overwhelming response bias towards one target or the other.

Figure 1 also shows the pooled labeling data, correct responses being indicated by the darkened portion of each histogram and errors by the white portion. It is obvious that subjects are identifying the intended productions at chance level: they cannot distinguish *cod* from *card*.

To determine whether subject's judgments were influenced by vocalic duration, regardless of what they had intended as speakers, we re-plotted the same data according to perceptual judgments. In Figure 2 *cod* judgments are plotted above the horizontal axis and *card* judgments below. If duration had influenced these judgments, the proportion of *cod* responses would have decreased, and the proportion of *card* judgments would have increased from left to right with increasing values of standard deviation. But no such correlation appears. Not even the positive and negative extremes of standard deviation are consistently labeled. Thus we have evidence that a distinction reliably made in production has no effect upon perception.

A possible explanation for this curious state of affairs is that, since the thirties, when the data were gathered on which Kenyon's and Thomas' discriptions were based, long and short /ɑ/ have begun to merge in this dialect. If such a linguistic change *were* in progress, we might indeed expect to find that habits of production persisted after a distinction had ceased to have any linguistic significance. This would mean merely that speakers were wasting effort in distinguishing words that had effectively become homophones. Note that the converse possibility—a linguistic distinction maintained in perception but unsupported in production--is unlikely, since it would result in misunderstandings.

The descriptions of Thomas and Kenyon, however, are based on impressionistic data, and we cannot be certain that the perceptual distinction existed even when the dialect was described. If it did not, then we cannot conclude that a change is in progress.

But whether a change is in progress or not, there is another way to interpret this phenomenon. The pronunciations of such words as *cod* and *card* may function to mark a dialectal rather than a lexical difference. It would be interesting to determine whether subjects could make a dialect judgment on the basis of these words if, and only if, they knew what lexical items were intended. We intend to pursue this question further.

REFERENCES

Kenyon, S. *American pronunciation* (10th ed.). Ann Arbor, Mich.: George Wahr, 1961.

Kurath, H. *A phonology and prosody of modern English*. Ann Arbor, Mich.: University of Michigan Press, 1964.

Labov, W., Yaeger, M., & Steiner, R. *A quantitative study of sound change in progress* (Report on Contract NSF-GS-3287). Philadelphia: University of

Pennsylvania, 1972.

Thomas, C. K. An introduction to the phonetics of American English (2nd ed.). New York: Ronald Press, 1958.

DECAY OF AUDITORY MEMORY IN VOWEL DISCRIMINATION*

Robert G. Crowder+

Abstract. Two experiments on same-different vowel discrimination are reported. In each the main variable was the duration of a silent delay between the two items being judged. As would be expected from the assumption that such judgments depend at least partly on auditory sensory memory, it was found that longer delays led to poorer discrimination than shorter delays. The auditory memory loss seems to be asymptotic at about three seconds, whether it is measured by correct discrimination or (as in one part of the second experiment) by the contextual influence of the first vowel on identification of the second.

INTRODUCTION

For all the work some of us have done on auditory sensory memory, we know very little about its time course. What evidence there is comes either from scattered reports using totally noncomparable methods or from experimental techniques that are not ideal for addressing the decay question. Still, most experts would probably agree that echoic memory does not remain available forever and that it decays slower than iconic memory. There have been two research programs that sought data relevant to the decay question, both using a form of masking to uncover properties of auditory memory:

In Massaro's experiments related to this topic, for example (see Massaro, 1970), a single tone selected from two possibilities is presented for a recognition response (high/low). The phenomenon of interest is that an unrelated masking tone presented just after the test stimulus impairs correct responding in a way that depends on the interval between target and mask. If the mask is delayed by about 250 msec, the response is unimpaired, but more immediate masks reduce performance considerably. It is the damage done by the mask that has led Massaro to infer the existence of auditory memory from this demonstration. In the stimulus suffix effect, discovered by Dallett (1965) and elaborated by Crowder and Morton (1969), the target memory trace is a hypothesized package of sound information about the last item in a memory-span

*Also Journal of Experimental Psychology: Human Learning and Memory, in press.

+Also at Yale University, New Haven CT.

Acknowledgment. I acknowledge with pleasure the assistance of Virginia Walters in this research. Thanks also to Bruno Repp for his comments on an earlier draft of this paper. This research was supported by NSF Grant No. BNS77-07062 to R. Crowder, and NICHD Grant HD01994 and BRS Grant RR05596 to Haskins Laboratories.

type list. Performance on this last item is badly damaged by an extra word, the stimulus suffix, presented as if it were the next item in the list. The suffix can be semantically unrelated to the rest of the list and need not be recalled. Again, auditory storage is inferred from the vulnerability of the target (last memory item in the list) to masking from the suffix.

In both the recognition masking of tones and the stimulus suffix paradigms, increasing the interval between the target and the mask leads to improved performance, up to a point. Massaro (1970) found this improvement reached asymptote at around 250 msec, and Crowder (1969) found that a suffix delayed by more than about 2 seconds had no effect on performance. Both these asymptotes were used as estimates of the duration of auditory memory (Crowder, 1969, page 261; Massaro, 1972, page 129). The reasoning was that masking would become ineffective when the target information in the sensory store had decayed. Although this claim by itself is quite true, it is invalid to conclude anything about decay from the time at which masking becomes no longer effective: When the mask is delayed, in these paradigms, the sensory trace might remain intact but meanwhile the subject has had the opportunity to encode the information in it; if the subject has been able to incorporate the information contained in the sensory trace to some more permanent format, then it makes no difference whether the mask does or does not destroy this information later.

Watkins and Todres (1980; see also Watkins & Watkins, 1980, Experiment 6) have recently reported several experiments on delayed suffixes. They offered evidence that the interval before a delayed suffix was indeed being used by subjects for readout of auditory information into some more permanent form of memory. They also confirmed Crowder's (1971, page 339) speculation that decay might be very much slower than originally conjectured by Crowder and Morton.1 Watkins and Todres have correctly observed that whereas the absence of a suffix effect after some delay says nothing about whether the auditory trace survives that long, the <u>presence</u> of a suffix effect at some delay does suggest the survival of the trace for at least that long. They found that if they prevented subjects from engaging in readout of the target information during the delay (between target and suffix), an appreciable suffix effect was obtained after 20 seconds.

Although the suffix experiment may thus be forced to yield acceptable inferences about minimal survival time of auditory memory, it is not ideal for this purpose. The portion of performance in the suffix experiment that is interesting for the analysis of auditory memory -- relative performance on the last serial position -- is superimposed on a background of highly complicated and strategy-prone short-term memory functions. For example, to demonstrate the 20-second-delayed suffix experiment, Watkins and Todres had to engage the subjects in a lively mental arithmetic task between the last memory item and the suffix item. We know that even so mundane a task as remembering a series of meaningless items in order engages several types of mechanisms -- grouping, cumulative rehearsal, efforts at semantic coding, articulatory loops, and so on -- and many of these mechanisms are quite likely to interact with serial position. Accordingly, it would be a boon to be able to study auditory memory and its decay properties in the context of a simpler task. That is the purpose of the research reported in this paper.

Pisoni (1973) has used a same-different speech discrimination task to study the decay of auditory memory (see also Repp, Healy, & Crowder, 1979; the background for these investigations is covered in Crowder, in press). In this task, the subject hears two speech sounds -- perhaps vowels similar to the /i/ and /I/ in BEET and BIT. The two sounds are typically quite close to each other acoustically, so that perhaps they both sound like one or the other of these two phonetic segments. The subject must decide whether the two are identical physically or not. Especially in the case where both items sound like only one of the possible phonetic segments, the reasoning is that auditory memory must have some role in correct performance. Consider the subject receiving the second of the two items to be judged: If the second item has the same name (phonetic label) as the first, then the only way the subject can tell whether they are physically identical is by remembering the sound of the first until the second arrives.

Pisoni (1973) set the delay between the two vowel stimuli at intervals from one-half to two seconds. He found that performance was poorer at the longer separations, as would be expected if the sound of the first item -- its auditory memory trace -- were decaying during the interval between stimuli. The logic of isolating sensory memory contributions through manipulation of delay in a successive discrimination task is not at all unconventional. Kinchla (1973) observes that such a task provides "...a rather direct approach to ´sensory memory´ processes..." In his experiment, subjects heard a compound tone and then, after a variable delay, had to make a fine intensity discrimination between a single probe tone and its corresponding element in the original compound; performance steadily decreased from compound-probe intervals of one-half to two seconds. Hanson (1977) found poorer performance in a "physical match" (same/different) task with interstimulus intervals of 570 msec than 250 msec, using stop-vowel CV syllables.

The present experiments were planned to test other intervals than those Pisoni used, in order to get an estimate of decay rate in auditory memory. This research cannot settle whether the auditory memory believed to support same-different speech discrimination is the same auditory memory that has been studied in the suffix experiments (*Precategorical Acoustic Storage*). That question needs a different kind of experiment. However, the same-different discrimination task is obviously a more direct and simple context in which to measure the auditory store and thus a useful context in which to ask about decay.

EXPERIMENT 1

Experiment 1 comprised two parts. In the first, there were 10 main conditions defined by 10 stimulus onset asynchronies separating the two items on each trial. These were set at 0, 200, 400, 600, 800, 1000, 1200, 1400, 1600, and 1800 msec. Since the vowels were 300 msec long, the first two of these conditions included physical overlap between the two items. It developed that at least one of the overlap situations was sharply inferior to the longer stimulus onset asynchrony conditions and subjects complained that they were confusing. Thus, after testing 20 subjects in the original design, we eliminated the two shortest stimulus onset asynchrony conditions and continued for another 20 subjects.

Method

Stimuli. The stimulus items were three-formant, steady-state, synthetic vowels similar to those used by Repp et al. (1979). These stimuli spanned the continuum from the vowel /i/ to /I/. The first formant center frequencies ranged from 269 to 397 Hz, the second from 2296 to 2030 Hz, and the third from 3019 to 2632 Hz, all in roughly logarithmic steps, for the continuum of eight. In this study, the fourth and fifth tokens were left out so as to enhance the contrast between within- and between-category decisions. The present set of vowels correspond to Stimuli 1, 2, 3, 6, 7, and 8 from Table 1 of Repp et al. (1979, page 139). The formant bandwidths were 63, 94, and 110 Hz, respectively. The vowels were 300 msec long and were produced on the Haskins Laboratories OVEIIIc synthesizer. Overall amplitude rose sharply over the first 50 msec, then remained uniform until a symmetric fall over the last 50 msec. Fundamental frequency declined gradually from 125 to 80 Hz throughout the utterances.

A different test tape was prepared for each of the 10 stimulus onset asynchrony conditions. On each, there were 18 pairs of identical tokens (1-1, 2-2, and so on, each repeated three times) where the correct answer was SAME. The other 42 pairs on each tape contained 16 "one step" DIFFERENT trials (1-2, 2-1, 2-3 and so on), 8 "two step" pairs, and 18 more widely spaced DIFFERENT pairs contrasting the /i/ items (1, 2, 3) with the /I/ items (6, 7, and 8). These 60 trial types were arranged on the tape in a different random order for each stimulus onset asynchrony.

Design and procedure. The subjects in Part One (10 different stimulus onset asynchrony conditions including 0 and 200 msec) received their tapes in an order determined by a balanced Latin square (complete control over first-order sequential effects). The subjects in Part Two followed the same Latin square design but the tapes with 0 and 200 msec stimulus onset asynchrony were simply deleted; thus they had 120 fewer trials than the first squad of subjects. Instructions were explicit about the experimental design and stressed that the criterion for a "same" response was to be *exact physical identity*.

Following each trial, there was a five second pause before the next trial. There were no warning sounds to mark trials or response periods. The subjects had a numbered answer sheet with the letters *s* and *d*, which they were supposed to circle, indicating their response for that trial. A practice tape consisting of 6 sample trials was presented after the instructions.

Subjects. The subjects were 40 college-age adults from the New Haven area, some Yale students serving as part of a course requirement and some volunteering to serve for pay.

Results and Discussion

The mean overall proportions correct for the two parts of Experiment 1 (SAME and DIFFERENT trials combined) are shown in the first two rows of Table 1 as a function of stimulus onset asynchrony. For this analysis, the two kinds of trials were not weighted (see the d' analysis, below). Two things are quite clear from inspection: There is some loss in discrimination as a function of delay, as we would expect from the Pisoni (1973) result. Secondly, the function is far from asymptotic over the range studied here.

Analysis of variance on these data confirmed the reliability of the basic delay effect. For this analysis the two parts were combined and only the data from stimulus onset asynchronies 400-1800 were included. To have included the 0-msec delay would have produced a misleadingly high $\underline{F}$ ratio because this condition was so extraordinarily poor relative to the others. For the combined data, the delay effect was highly significant statistically, $\underline{F}$ (7,273) = 5.88, MSe = 7.90, $\underline{p}$ <.001.

It can be objected that these data may be influenced to some unknown degree by changes in response criteria for judging two items physically identical, across the different delay conditions. Such arguments (see Macmillan, Kaplan, & Creelman, 1977) make the case for analyzing the data in terms of Statistical Decision Theory. Tables have recently become available for transforming variable-standard same/different data into d′ (Kaplan, Macmillan, & Creelman, 1978).[2] The task is conceived as one where the subject is set to "detect sameness" and one uses as false alarms the proportion of SAME responses when the two items were in fact different. The data relevant to this analysis are shown in the second two rows of Table 1.[3] The conclusions of the conventional analyses are completely sustained by this unbiased analysis of sensitivity. Analysis of variance based on the 10 supersubjects from both parts of the experiment on the conditions in common (stimulus onset asynchrony 400 through 1800 msec) confirmed the reliability of the delay effect, $\underline{F}$ (7,63) = 3.92, MSe = .162, $\underline{p}$ < .05. Thus, no changing criterion for "sameness" across different stimulus onset asynchrony values can be held responsible for the declining performance observed here. Note that althougn the bias is not changing over intervals in a way that produces the decay effect, there is an overall strong bias in responding: This is indicated by the large d′ values and the relatively low (about 70%) rates of correct responding. The overall probability of saying SAME when the two stimuli were identical was very high, .919, and the corresponding rate of false alarms, SAME given different, was .378. This same bias was observed in the second experiment. To repeat, the important consideration is that a changing bias cannot account for the result of interest here.

Although the majority of trials in this experiment contained, by design, items from the same phonetic category, there were enough between-category pairs to inspect for a difference between the size of the decay effect in between- and within-category trials. This was done using stimulus pairs as the sampling variable. For each of the 12 within-category pairs where the correct response was "different," the number of errors made by all subjects on stimulus onset asynchronies 400-1000 and 1200-1800 was tallied separately. The reliability of the decay effect for the within-category data was verified by a paired t-test, $\underline{t}$ (11) = 2.83, $\underline{p}$ < .01. The same was done for the 18 between-category pairs and again the decay effect was reliable, $\underline{t}$ (17) = 6.43, $\underline{p}$ < .005. Going by the size of the $\underline{t}$ values, one might suppose the effect was larger for the between-category pairs and indeed the raw differences between the short- and long-delay conditions were significantly larger for the between-category pairs than for the within-category pairs, $\underline{t}$ (28) = 3.00, $\underline{p}$ < .005. However, the between-category pairs all spanned a larger physical distance than the within-category pairs and so there were many fewer errors in

TABLE 1

Discrimination Performance in Experiment 1

Stimulus Onset Asynchrony (milliseconds)	Measure			
	Proportion Correct		d′	
	Part 1	Part 2	Part 1	Part 2
0	.497	---	1.01	---
200	.729	---	2.86	---
400	.727	.728	3.21	2.94
600	.707	.729	3.17	3.51
800	.747	.709	3.48	3.01
1000	.721	.713	3.02	3.25
1200	.718	.713	3.17	3.12
1400	.696	.694	2.64	2.74
1600	.679	.694	2.79	2.77
1800	.682	.684	2.70	2.71

the former. If one wishes to take the ratio of the difference between the short (S) and long (L) intervals relative to the total errors made for a pair -- (L - S)/(L + S) -- the difference is reversed. By this latter measure, there was a significantly larger delay effect in the between-category data than in the within-category data, _t_ (28) = 5.40, _p_ < .005. The conclusion has to be that delay does not have a larger effect on within-category pairs than on between-category pairs. This was the outcome of the Pisoni (1973) and Repp et al. (1979) studies, too.

The component of performance in discrimination that can be assigned to auditory memory has quite plainly not reached asymptote by the longest interval tested in Experiment 1. The main purpose of Experiment 2 was to expand the range of intervals tested.

EXPERIMENT 2

The stimulus onset asynchrony values used in this second study were 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000 msec. In most other respects the experiment was similar to Experiment 1 except for one additional feature: Experiment 2 also included a complete run through the materials for each subject in which identification, rather than discrimination, was measured. Repp et al. (1979) had found that the items within a pair exerted roughly symmetrical, contrastive effects on labeling. That is, they observed that when members of a pair were being labeled phonetically as /i, I, or ɛ/, the identity of the other pair member influenced the item being labeled. The effect was contrastive, which means that if an ambiguous vowel between /i/ and /I/ were presented, hearing it in the context of an unambiguous /i/ made it sound more like /I/. That the effect was symmetrical means that the first item in the pair influenced the second about as much as the other way around. Repp et al. suggested that these context effects on phonetic labeling were produced by mechanisms within auditory memory because in conditions where auditory memory was removed by delay or by masking, little contextual influence was found.

By analogy, the contrastive effects on phonetic labeling can be compared with visual brightness contrast: A given shade of gray appears brighter if it occurs in the context of a dark background than if it appears in a light background. For any successive contrast to work, it might be suggested that the two items would have to reside together in memory. If so, then we can understand why Repp et al. found less contrast when they compromised the auditory storage of items during the interstimulus interval. It follows that contrast effects could be used as an independent measure of the duration of auditory memory.

A word should be added about what _causes_ contrastive context effects: The generalization of importance is that one vowel affects the label applied to another provided they are different and provided they occupy auditory memory together. In recent publications (Crowder, 1978, in press) I have begun to advance a theory that covers these findings. The central assumption relevant to context effects (Crowder, in press) is that auditory-memory representations interact by _frequency-specific_ inhibition of each other. That is, if auditory memory representations of two items occur close together in time, and on the same channel, they will tend to inhibit each other and this inhibition will be greatest in spectral regions where they contain overlapping energy. If two vowels are similar except for the placement of one or two formants, this frequency-specific inhibition will produce contrast: The formants associated with the vowels used here have very considerable overlap relative to their center frequency differences. This means that two vowels' formants will have an area of intersection and also each will have an area not in common with the other. If inhibition between them is frequency specific, the intersection in the vowels' formants will suffer the most, leaving the non-intersecting formant area in each vowel relatively intact. Since the non-intersecting regions were what made the two vowels distinctive in the first place, eliminating the region in common will enhance their distinctiveness mutually, and will lead to contrastive identification. See Crowder (in press) for further explanation. This interpretation is consistent with a theory that applies equally well to the suffix and vowel-discrimination tasks and covers essentially all known evidence on the suffix effect (Crowder, 1978).

Method

Stimuli. A different set of vowels was used in Experiment 2. This was primarily in order to increase the generality of the research program. The 13-item continuum used in this study crossed the vowel space in such a path as to include approximate prototypes of /ɑ/, /ʌ/, and /æ/, which correspond to the vowel sounds in COT, CUT, and CAT, respectively. To achieve this, the formant frequencies shown in Table 2 were set on the OVEIIIc synthesizer. Included in Table 2 are the overall identification data when each of the thirteen tokens was presented with itself -- that is, on SAME trials -- collapsed over inter-item delays. These data show that the subjects were quite willing to accept this as a three-vowel continuum. In other respects, the stimulus items were similar to those of Experiment 1.

Each test tape contained 34 pairs, of which 13 were SAME trials (1-1, 2-2,...,13-13), 11 were two-step DIFFERENT trials (1-3, 2-4,...,11-13), and 10 were three-step DIFFERENT trials (1-4, 2-5,...,10-13). It was arbitrarily decided to use only DIFFERENT trials that *ascended* in terms of the numbering of Table 2 (that is 1-4, but not 4-1). These 34 pair types were randomly ordered 10 times and placed on tapes otherwise differing only in the stimulus onset asynchrony -- 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000 msec. The interval between trials was 4 seconds.

Design and procedure. Every subject went through the 10 tapes twice, first in an identification experiment and second in a same/different discrimination experiment. In the former, they were instructed to listen carefully to the *second* stimulus in each pair and to identify it by circling one of the words (COT, CUT, or CAT) on a numbered answer blank. It was expected that the *first* item in each pair would provide a contextual influence on this labeling, to the extent the two items occupied auditory memory together. The 10 tapes were presented in a balanced Latin square order.

In the second part of the experiment, the same 10 tapes were presented to each subject in the reverse order to that used in the first part. Here, the instructions were to make a same/different judgment for each pair based on the same criteria explained in the previous experiment. Again, a practice tape

TABLE 2

Stimuli used in Experiment 2

		Formant Structure			Labels on SAME Trials		
Stimulus Number		F_1	F_2	F_3	/ɑ/	/ʌ/	/æ/
1	/ɑ/	728	1091	2431	.969	.015	.015
2		713	1107	2431	.964	.031	.005
3		702	1123	2431	.967	.023	.010
4		687	1139	2431	.918	.072	.011
5		668	1156	2431	.667	.323	.010
6		653	1172	2396	.536	.459	.005
7	/ʌ/	639	1189	2396	.182	.815	.003
8		644	1279	2396	.026	.933	.041
9		644	1364	2396	.023	.795	.182
10		649	1456	2396	.010	.436	.554
11		653	1543	2413	.005	.221	.774
12		658	1635	2413	.003	.038	.969
13	/æ/	658	1719	2413	.003	.005	.990

was used to provide familiarity with the sounds.

Subjects. The subjects were 40 young adults from the same source as in Experiment 1.

Results and Discussion: Discrimination

The discrimination results are given in Figure 1, which shows the overall proportion of correct same/different judgments as a function of stimulus onset asynchrony. As in the first experiment, performance began to drop sharply between one and two seconds. However, the figure shows little change after three seconds, suggesting that auditory memory -- to the extent it represents a decaying source of information for same/different responding -- has been lost by three seconds.

The same picture is provided by the d′ analysis shown in Figure 2. If anything the results are cleaner when corrected this way for possible criterion artifacts. Statistical analysis confirmed the reliability of the findings in Figures 1 and 2. Separate analyses of the untransformed error types "same" on DIFFERENT trials and "different" on SAME trials showed that each component of the pooled errors in Figure 1 was statistically significant, *F*′s (9,351) = 4.82 and 6.49, respectively, (MSe′s = 4282.1, 1860.1), *p*′ s < .0001 Analysis of variance on d′s again used supersubjects of four individuals each. There were 10 of these supersubjects and the d′ variance associated with stimulus onset asynchrony was highly significant, *F* (9,81) = 7.55, MSe = 2163.74, *p* < .001. As in Experiment 1, there was no evidence that the delay was more potent for the within- than for the between-category pairs: In this study, the identification results provided only five pairs that could convincingly be called within-category (1-3, 1-4, 2-4, 7-9, and 11-13 -- see Table 2). One of these showed reduced errors from the short- to the long-interval conditions while the other four showed increased errors. The between-category pairs showed reliable and consistent delay effects, however, *t* (15) = 3.78, *p* < .005. As before, the auditory component was not by any means restricted to the cases where items betng discriminated match in phonetic category.

Performance remained quite good even after the component being attributed here to auditory memory had decayed to asymptote. However, not too much importance should be attached to the specific levels of correct responding. These reflect, among other things, the mixture of easy, three-step discriminations (where performance ranged from .875 to .825) and the more difficult two-step discriminations (.670 to .580). Furthermore, there was a strong bias for responding "same," as is evident in the correct "same" responses on trials where the two items were identical, where hits ranged from .960 in the 500-msec stimulus onset asynchrony condition to .875 in the 4000-msec condition. Corresponding "same" responses on DIFFERENT trials ranged from .235 in the 500-msec condition to .312 in the 3000-msec condition. The mean proportions in Figure 1 thus represent none of the exact performance levels obtained. The important thing of course is the regularity of the data and not absolute levels of accuracy.

Correct performance was also influenced by the particular items being discriminated along the continuum from /ɑ/, /ʌ/, through /æ/. Table 3

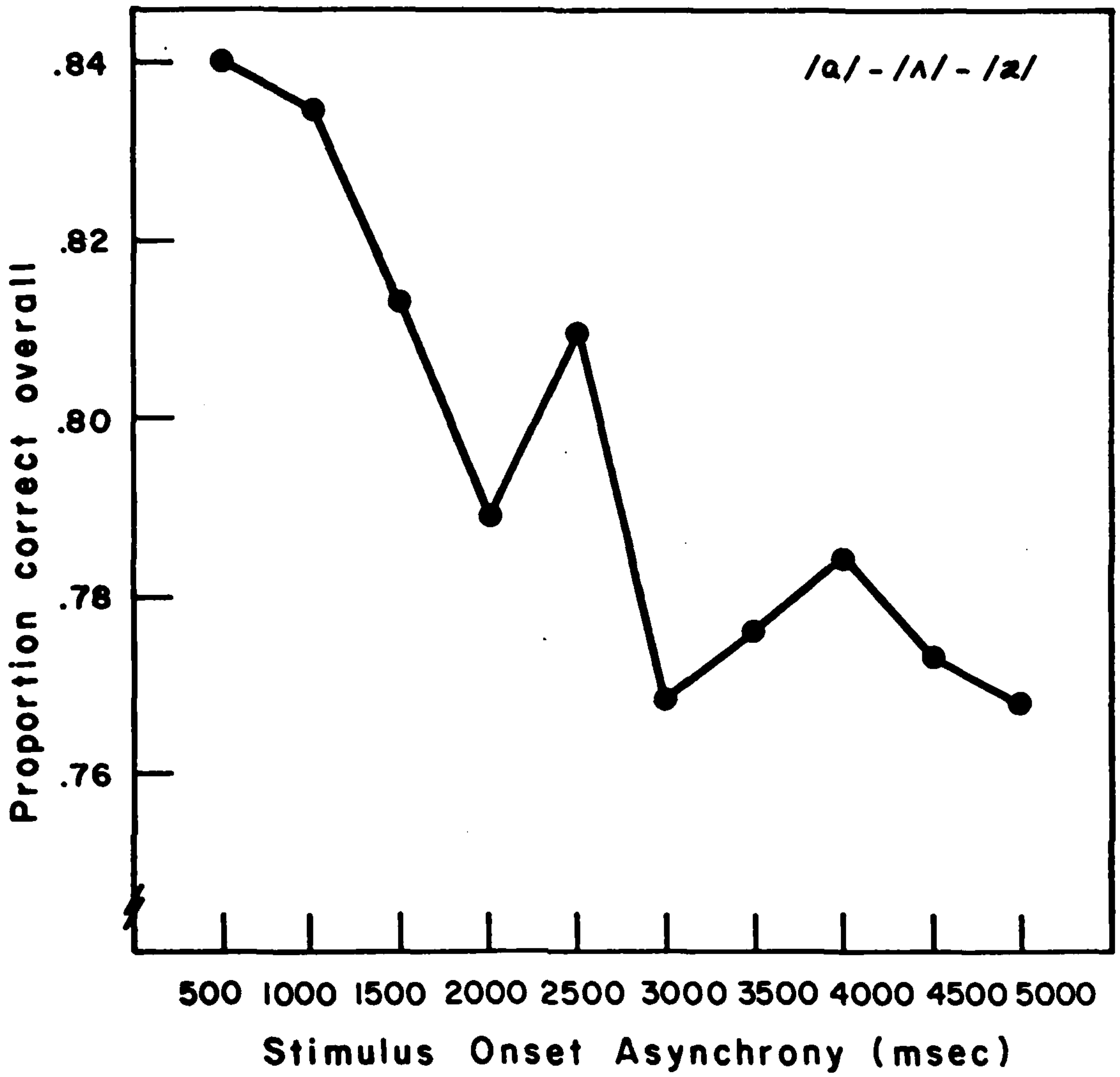


Figure 1. Proportion of correct responses, overall, in Experiment 2.

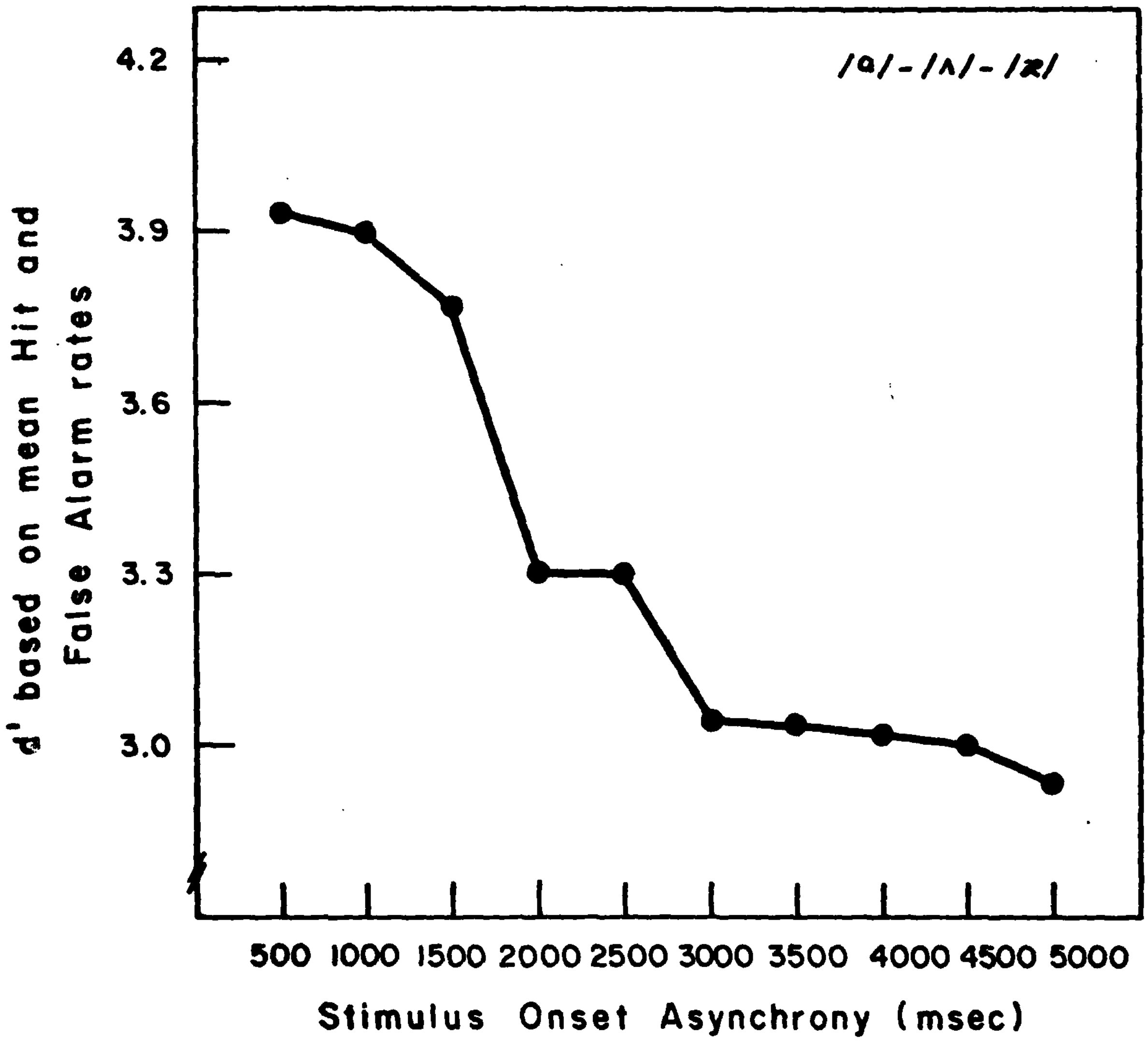


Figure 2. Same/different discrimination sensitivity (d') as a function of stimulus onset asynchrony in Experiment 2. Each point represents performance of ten supersubjects based on four individuals apiece.

TABLE 3

Proportion correct same/different discrimination (combined SOAs)

		DIFFERENT			
SAME		Two-Step		Three-Step	
Pair	Proportion	Pair	Proportion	Pair	Proportion
1- 1	.945	1- 3	.140	1- 4	.373
2- 2	.947	2- 4	.167	2- 5	.657
3- 3	.940	3- 5	.485	3- 6	.735
4- 4	.937	4- 6	.515	4- 7	.747
5- 5	.870	5- 7	.490	5- 8	.883
6- 6	.880	6- 8	.767	6- 9	.957
7- 7	.855	7- 9	.885	7-10	.987
8- 8	.920	8-10	.843	8-11	.973
9- 9	.917	9-11	.825	9-12	.975
10-10	.917	10-12	.887	10-13	.975
11-11	.897	11-13	.885		
12-12	.910				
13-13	.957				

shows the proportion correct overall for each of the SAME and DIFFERENT pairs used in the experiment. Quite clearly, the /æ/-end of the continuum was easier than the /ɑ/-end. These differences reflect no doubt the spacing of tokens shown in Table 2. However, the important question is whether the main decay results were general across these stimuli, which differed widely otherwise in discrimination difficulty. The answer is reassuring: Among the 13 types of SAME trials (1-1, 2-2, and so on) performance at the shortest interval was better than performance at the longest interval in ten cases, with one tie and two reversals, *p* = .019 by a sign test. Among the 21 DIFFERENT trial types, there were 17 pairs showing the same difference, with one tie and three reversals, *p* = .006 by a sign test. Thus the extreme variability in pair difficulty is another reason for skepticism about the absolute values of the means shown in Figures 1 and 2 but it does not discount the generality of the time profile shown there.

One might very well wonder whether the group asymptote of 3000 msecs is representative of the performance of many individual subjects. The analyses of variance reported here insures that the decay effect generalized across variability due to subjects and evidence has been presented, above, for such generality across items. But the generality of the asymptote requires stronger arguments. There are not enough data for each subject to calculate individual regressions of performance on delay. However, the d′ values for the ten supersubjects could be inspected across the ten delays for that purpose. As a rough estimate of where these ten functions reached asymptote, the interval with the lowest d′ was determined. For one supersubject, this minimum was at 500 msec stimulus onset asynchrony, for another, it was at 5000, and for two each of the remaining eight, it fell at 3000, 3500, 4000, and 4500 msec. This near rectangular distribution of the minima is consistent with the generalization that performance does not change after 3000 msec.

Results and Discussion: Identification

The identification results from SAME trials have already been displayed in Table 2. These data are collapsed over stimulus onset asynchrony but, as will be seen presently, stimulus onset asynchrony did not matter for the SAME trials. The identification data of Table 2 show there were two boundaries -- that between /ɑ/ and /ʌ/ falling between stimuli 6 and 7 and the one between /ʌ/ and /æ/ falling between stimuli 9 and 10. The question is now whether these boundaries shifted when subjects were identifying the exact same tokens but in the context of a prior item from "higher up" on the numbered continuum of Table 2 (recall that the prior context always came from this direction). To replicate the Repp et al. finding of contrast, the present results would have to show that a given token sounded as though it came from "lower down" on the continuum if it occurred on a DIFFERENT trial than if it came on a SAME trial. In terms of boundary locations, this means the boundaries would shift to the opposite direction -- to a smaller numerical value.

The data relevant to this point are shown in Figure 3, which gives a summary of context effects. Here, the data of Table 2 on identification are broken down into the different stimulus onset asynchrony conditions -- grouped by two′s for stability. Boundaries were estimated by linear regression on stimuli 3-8 for the /ɑ/-/ʌ/ transition and on stimuli 8-12 for the /ʌ/-/æ/

transition.[4] The two boundaries associated with the three phonetic segments are collapsed in such a way that the numerical boundary measures on the vertical axis show the mean stimulus number of the two boundaries.

Figure 3 shows clearly that for SAME trials, the stimulus onset asynchrony made no difference. However, on DIFFERENT trials, boundary locations shifted in the expected direction -- toward lower numerical values -- when there had been a recent context item. If the SOA was longer than three seconds, it was as if there had been no context at all, but at shorter intervals, context changed the labels applied to the second member of the pair. The convergence of phonetic labeling on SAME and DIFFERENT trials at three seconds is consistent with the suggestion that contrast operates when the two items in question occupy auditory memory together. The particular time interval at which these data converge is in approximate agreement with the estimate of asymptotic decay that was based on discrimination, reported above.

Statistical analyses confirmed the reliability of the picture presented in Figure 3. For the short stimulus onset asynchronies combined (500 msec through and including 2500 msec), 26 out of 37 nontied subjects placed stimuli 6 and 9[5] farther down the numbered continuum on DIFFERENT trials than on SAME trials, _p_ = .01. The context effect was surprisingly general across stimuli as well as across subjects: For the short and long intervals, as defined above, each of the 11 stimuli labeled in a DIFFERENT context (numbers 3, 4,...,13) was given a mean "placement score" along the continuum. This placement was simply a weighted average of the three phonetic labels assigned by subjects.[6] The same placement score was available from the SAME trials.

The question was whether a given stimulus item would receive lower (that is, farther down the list) placement on the DIFFERENT trials than on the SAME trials. At the short intervals, this was the result for 9 of the 11 items, _p_ = .033. Furthermore, 10 of the 11 items also showed the full pattern of Figure 3 -- a bigger directional difference between SAME and DIFFERENT trials at the short than at the long intervals, _p_ = .006. Thus, the contrastive context effects on labeling generalize both across subjects and across individual vowel tokens.

It is somewhat surprising that the context effects proved so consistent across stimuli. One would have expected primarily the ambiguous items to show influence of context. Therefore, further analyses were undertaken to examine the relation between the degree of the context effect and the position of a stimulus on the continuum. For this purpose, only the short (500 to 2500 msec) stimulus onset asynchrony data were used. For each of the 11 stimuli that were labeled on DIFFERENT trials, two placement scores were compared, one on SAME trials and one on DIFFERENT trials. A positive difference means the vowel in question showed different phonetic labeling, in the predicted direction, when it followed another vowel from the continuum. These differences in placement are shown in Figure 4 in arbitrary numbers that reflect the calculation of placement scores. The figure makes obvious that, although all but two items showed a "positive" context effect, as reported above, the size of that context effect was related in an interpretable fashion

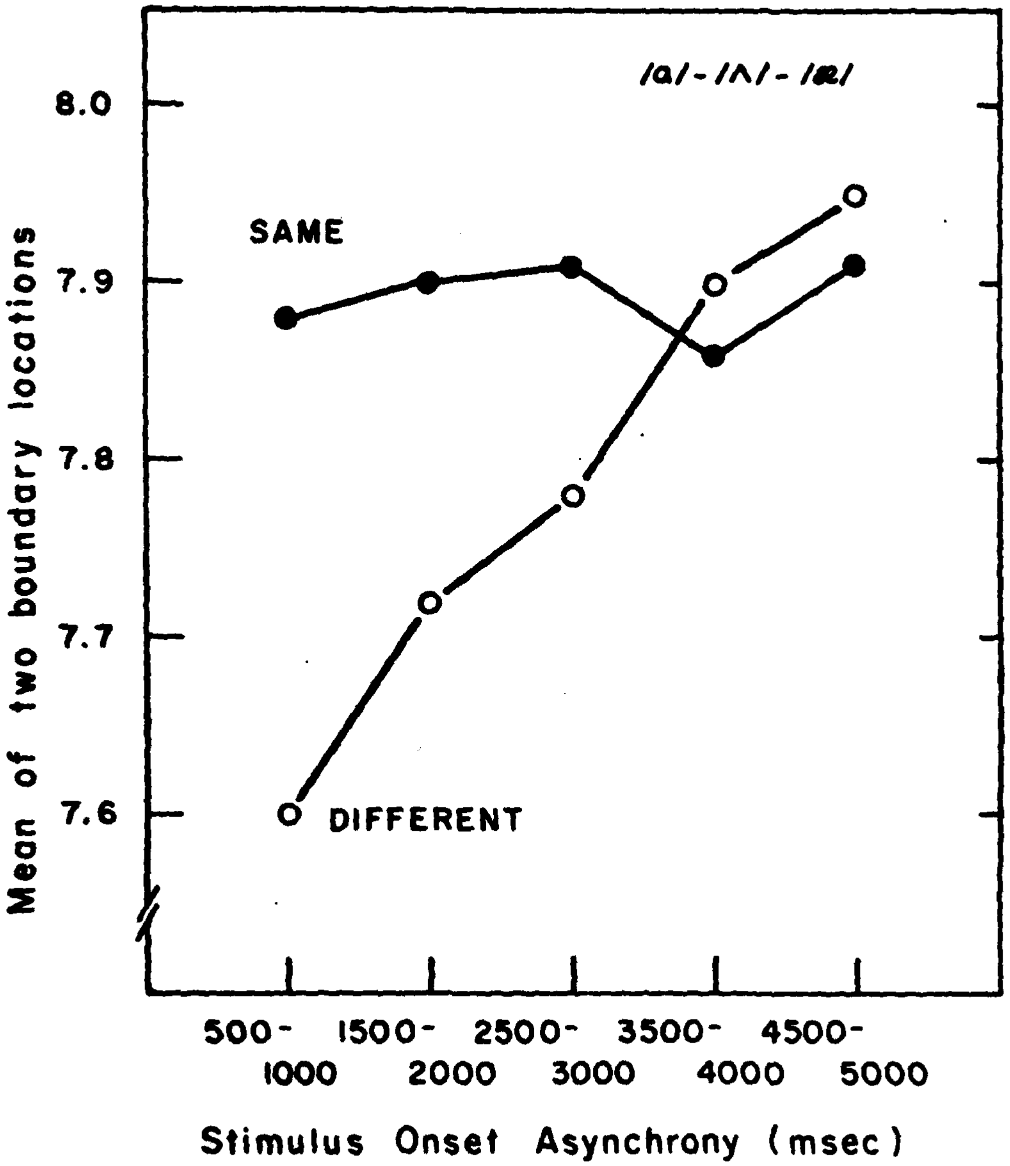


Figure 3. The relation between boundary placement and stimulus onset asynchrony for SAME and DIFFERENT trials in the phonetic identification phase of Experiment 2. The numbers on the vertical axis represent the mean of the two boundary values.

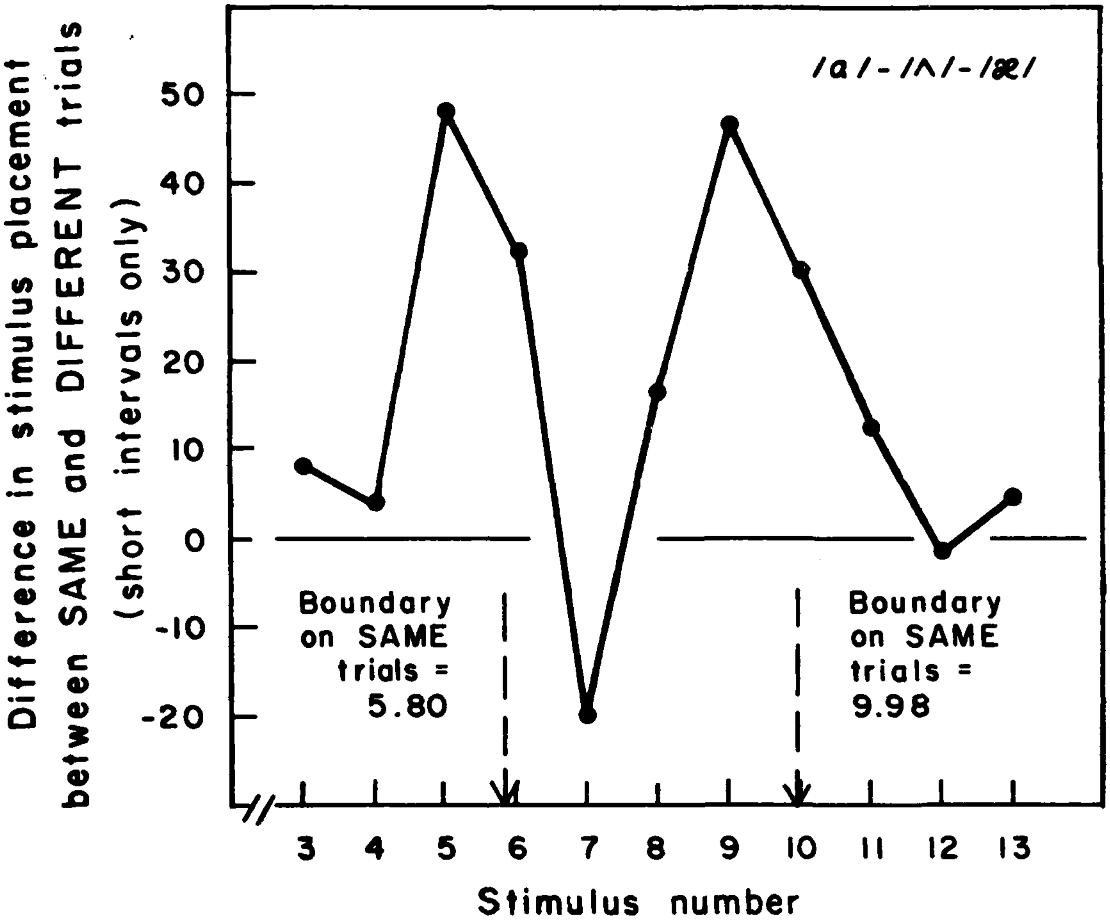


Figure 4. The relation between stimulus number and the size of context effects on labeling. A high positive score means a particular vowel was labeled as coming from farther away from its prior context vowel than it would have been if that prior context had been the same vowel itself. The arrows show combined category boundaries for SAME trials.

to the category boundaries derived from SAME trials. There were two peaks in the contextual influence and they coincide closely with the two category boundaries. In other words, as one might expect, it was the ambiguous items that were most susceptible to context.

GENERAL DISCUSSION

The main goal of these studies was to provide parametric data on the decay of auditory sensory memory. The results give a consistent estimate that this decay is asymptotic at close to three seconds, for the successive discrimination task used here. The phonetic labeling data of Figure 3 show another manifestation of auditory memory -- context influence on identification -- and this influence disappears at just the same time.

Experiments using related techniques to investigate memory for tones (for example, Harris, 1952; Moss, Myers, & Filmore, 1970) do not necessarily converge on the same estimate; however, there are typically not enough intervals studied in these experiments to establish an asymptote, and, even if there were, the stimuli and tasks are different enough to discourage comparison. On the other hand, the estimate of three seconds is close to the value suggested by Crowder and Morton (1969), even though that estimate was only a shot in the dark.

Although the high performance levels in these experiments demonstrate that other factors besides transient auditory memory support performance in this task setting, it is a relatively uncomplicated task compared to the suffix experiment. If further research suggests that the successive vowel discrimination task used here taps the same auditory memory store that has been so extensively studied in the suffix experiment, it may be advisable to focus on the former rather than the latter in future work because it is so much more direct a method. Perhaps the least encouraging evidence on this point is the finding of Watkins and Todres (1980) and of Watkins and Watkins (1980) that suffix-like effects occur following filled delays of up to 20 seconds. It will be for further research to clarify what are the boundary conditions on this delayed suffix effect and to establish whether it has the same functional properties as the immediate suffix effect, such as sensitivity to phonetic class and to physical source channel.

The most intuitively plausible model for how auditory memory is used in speech discrimination is that subjects try first to make a same/different decision based on phonetic labels and, only after that has failed, go on to consult auditory memory. The rule is "If the two sounds have different names, say 'different,' otherwise compare the sounds themselves." This model (see Crowder, in press, and Pisoni, 1973, for details) is apparently wrong. It anticipates that effects owing to auditory memory would be stronger in the within-category discriminations than in the between-category discriminations. Neither the present studies, the results of Pisoni (1973), nor those of Repp et al. (1979) gave evidence for the predicted interaction.

Perhaps subjects adopt some private categorical discrimination that does not match the conventional phonetic categories but nonetheless serves a similar role in performance on the within-category pairs. After listening to the items in the stimulus ensemble for some time, subjects might very well compare the two sounds. In that case, with "functional categories" constructed for the nominal within-category pairs, it would not be so surprising that there was about the same auditory influence in the within- and between-category judgments.

Does the three-second estimate from this research suggest any functional role for auditory memory outside the narrow task confines of this procedure? Of course, there are now only the most preliminary efforts to connect laws of information processing to real-time language processing. However, Stevens (1978, page 14) has noted the relationship between sentence-length utterances and breathing. He observes a close relation between syntactic structures and the pauses introduced by a speaker for the inspiration of breath. As Stevens notes, exigencies of breathing limit sentences, or other major syntactic structures, to a length of not more than two or three seconds. Thus, the three-second figure is of some linguistic interest in a way that could be related to speech production or comprehension. But this comment is no more than suggestive: For one thing, the echoic decay estimate comes from a situation where the trace is held in complete silence whereas the two- to three-second limit associated with the breath group is typically filled with speech.

REFERENCES

Crowder, R. G. Improved recall for digits with delayed recall cues. *Journal of Experimental Psychology*, 1969, *82*, 258-262.

Crowder, R. G. Waiting for the stimulus suffix: Decay, delay, rhythm, and readout in immediate memory. *Quarterly Journal of Experimental Psychology*, 1971, *10*, 587-596.

Crowder, R. G. Sensory memory systems. In E. C. Carterette & M. P. Friedman (Eds.) *Handbook of perception, Volume 9*. New York: Academic Press, 1978.

Crowder, R. G. The role of auditory memory in speech perception and discrimination. In T. Myers, J. Laver, & J. Anderson (Eds.) *The cognitive representation of speech*. Amsterdam: North-Holland Publishing Co., in press.

Crowder, R. G. & Morton, J. Precategorical acoustic storage (PAS). *Perception & Psychophysics*, 1969, *5*, 365-373.

Dallett, K. "Primary memory": The effects of redundancy upon digit repetition. *Psychonomic Science*, 1965, *3*, 237-238.

Hanson, V. L. Within-category discriminations in speech perception. *Perception & Psychophysics*, 1977, *21*, 423-430.

Harris, J. D. The decline of pitch discrimination with time. *Journal of Experimental Psychology*, 1952, *43*, 96-99.

Kaplan, H. L., Macmillan, N. A., & Creelman, C. D. Tables of d′ for variable-standard discrimination paradigms. *Behavior Research Methods and Instrumentation*, 1978, *10*, 796-813.

Kinchla, R. A. Selective processes in sensory memory: A probe comparison procedure. In S. Kornblum (Ed.) *Attention and performance IV*. New York: Academic Press, 1973.

Macmillan, N. A., Kaplan, H. I., & Creelman, C. D. The psychophysics of categorical perception. *Psychological Review*, 1977, *84*, 452-471.

Massaro, D. W. Preperceptual auditory images. *Journal of Experimental Psychology*, 1970, *85*, 411-417.

Massaro, D. M. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 1972, *79*, 124-145.

Moss, S. M., Myers, J. L., & Filmore, T. Short-term recognition memory of tones. *Perception & Psychophysics*, 1970, *7*, 369-373.

Pisoni, D. B. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, 1973, *13*, 253-260.

Repp, B., Healy, A. F., & Crowder, R. G. Categories and context in the perception of isolated steady-state vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, *5*, 129-143.

Sorkin, R. D. Extension of the theory of signal detectability to matching procedures in psychoacoustics. *Journal of the Acoustical Society of America*, 1962, *43*, 1745-1751.

Stevens, K. N. The speech signal. In J. F. Kavanagh & W. Strange (Eds.), *Speech and language in the laboratory, school, & clinic*. Cambridge, MS: MIT Press, 1978.

Watkins, M. J., & Todres, A. K. Suffix effects manifest and concealed: Further evidence for a 20-second echo. *Journal of Verbal Learning and Verbal Behavior*, 1980, *19*, 46-53.

Watkins, O. C., & Watkins, M. J. The modality effect and echoic persistence. *Journal of Experimental Psychology: General*, 1980, *109*, 251-278.

FOOTNOTES

[1] Our statement (Crowder & Morton, 1969, page 366) was that a store lasting at least on the order of a "few seconds" would be adequate for the functional role we had proposed for auditory memory.

[2] Sorkin (1962) has shown why a straightforward application of standard d′ tables to the same/different situation is inappropriate.

[3] For purposes of getting d′ values, subjects were combined into "supersubjects" of n=4 because many individual hit rates were close to or at 1.00. The mean data look essentially the same whether overall hit and false alarm rates are taken before calculation of d′ or these rates are calculated for each supersubject. For the purpose of statistical tests on d′ values, however, it is convenient to set up the supersubjects first.

[4] A check on the data of Table 2 will verify that performance on labeling within these ranges was unambiguously linear for the group data.

[5] These two stimuli, 6 and 9, were chosen because they represent performance from items just prior to the two respective boundaries on the identification test and therefore should represent ambiguous stimuli, especially subject to context effects.

[6] Specifically, the three identification responses /ɑ , ʌ , æ/ were assigned the numbers 1, 2, and 3, respectively. The total response to a stimulus for a given subject could then be characterized as an average of the numbers assigned to it. These averages were then compressed to a range from .33 to 1.00 for analysis.

THE EMERGENCE OF PHONETIC STRUCTURE*

Michael Studdert-Kennedy+

Abstract. To explain the unique efficiency of speech as an acoustic carrier of linguistic information and to resolve the paradox that units corresponding to phonetic segments are not to be found in the signal, consonants and vowels were said to be "encoded" into syllabic units. This approach stimulated a decade of research into the nature of the speech code and of its presumably specialized perceptual decoding mechanisms, but began to lose force as its implicit circularity became apparent. An alternative resolution of the paradox proposes that the signal carries no message: it carries information concerning its source. The message, that is, the phonetic structure, emerges from the peculiar relation between the source and the listener, as a human and as a speaker of a particular language. This approach, like its predecessor and like much recent work in child phonology and phonetic theory, takes the study of speech to be a promising entry into the biology of language.

The earliest claim for the special status of speech as an acoustic signal sprang from the difficulty of devising an effective alternative code to use in reading machines for the blind. Many years of sporadic, occasionally concentrated effort have still yielded no acoustic system by which blind (or sighted) users can follow a text much more quickly than the 35 words a minute of skilled Morse code operators. Given the very high rates at which we handle an optical transform of language, in reading and writing, this failure with acoustic codes is particularly striking. Evidently, the advantage of speech lies not in the modality itself, but in the particular way it exploits the modality. What acoustic properties set speech in this privileged relation to language?

The concept of "encodedness" was an early attempt to answer this question (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Liberman and his colleagues embraced the paradox that, although speech carries a linguistic message, units corresponding to those of the message are not to be found in the signal. They proposed that speech should be viewed not as a cipher on linguistic structure, offering the listener a signal isomorphic, unit for unit, with the message, but as a code. The code collapsed the phonemic segments (consonants and vowels) into acoustic syllables, so that cues to the

*Also in Cognition, 1981, 10, 301-306.

+Also Queens College and Graduate Center, City University of New York.

Acknowledgment. I thank Alvin Liberman, Ignatius Mattingly, and Bruno Repp for much fruitful discussion and advice. Preparation of the paper was supported in part by NICHD Grant HD 01994 to Haskins Laboratories.

component segments were subtly interleaved. The function of the code was to finesse the limited temporal resolving power of the ear. We typically speak and comfortably understand speech at a rate of 10-15 pnonomes/second, close to the rate at which discrete elements merge into a buzz. By packaging consonants and vowels into syllabic units, the argument went, we reduce this rate by a factor of two or three and so bring the signal within the resolving range of the ear.

This complex code called for specialized decoding mechanisms. More than a decade of research was devoted to establishing the existence of a specialized phonetic decoding device in the left cerebral hemisphere and to isolating the perceptual stages by which the supposed device analyzed the syllable into its phonetic components. This information-processing approach to speech perception exploited a variety of experimental paradigms that had seemed valuable in visual research (see Darwin, 1976, and Studdert-Kennedy, 1976, 1980, for reviews), but led eventually to a dead end, as it gradually became apparent that the undertaking was mired in tautology. A prime example was the proposal to "explain" sensitivity to features, whether phonetic or acoustic, as due to feature-detecting devices, and to look for evidence of such mechanisms in infants.

Current research has drawn back and is now moving along two different, though not necessarily divergent paths. The first bypasses the problems of segmental phonetic perception and focuses on what some believe to be the more realistic problem of describing the contributions of prosody, syntax, and pragmatics to understanding speech. The second path, with which I am concerned, reverses the procedure of the earlier encoding approach. Instead of assuming that linguistic units should somehow be represented as segments in the signal and then attempting to circumvent the paradox of their absence by tailoring a perceptual mechanism for their extraction, the new approach simply asks: What information does the speech signal, in fact, convey? If we could answer this question, we might be in a position not to assume and impose linguistic structure, but to describe how it emerges.

Consider the lexicon of an average middle-class American child of six years. The child has a lexicon of about 13,000 words (Miller, 1977), most of them learned over the previous four years at a rate of 7 or 8 a day. What makes this feat possible? Of course, the child must want to talk, and the meanings of the words she learns must match her experience: cat and funny, say, are more likely to be remembered than trepan and surd. But logically prior to the meaning of a word is its physical manifestation as a unit of neuromuscular action in the speaker and as an auditory event in the listener. Since the listening child readily becomes a speaker, even of words that she does not understand, the sound of a word must, at the very least, carry information on how to speak it. More exactly, the sound reflects a pattern of changes in laryngeal posture and in the supralaryngeal cavities of the vocal tract. The minimal endowment of the child is therefore a capacity to reproduce a functionally equivalent motor pattern with her own apparatus. What properties of the speech signal guide the child's reproduction?

We do not know the answer to this question. We do not even know the appropriate dimensions of description. But several lines of evidence suggest that the properties may be more dynamic and more abstract than customary descriptions of spectral sections and spectral change. For example, some half dozen studies have demonstrated "trading relations" among acoustically incommensurate portions of the signal (e.g., Liberman & Pisoni, 1977; Repp, Liberman, Eccardt, & Pesetsky, 1978; Fitch, Halwes, Erickson, & Liberman, 1980). Perhaps the most familiar example is the relation between onset frequency of first formant transition and delay in voicing at the onset of a stop consonant-vowel syllable: reciprocal variations in spectral structure and duration of delay produce equivalent phonetic percepts (Summerfield & Haggard, 1977). Presumably, the grounds of this and other such equivalences lie in the articulatory dynamics of natural speech, of which we do not yet have an adequate account. (For a review of studies of this type, see Repp, 1981).

A second line of evidence comes from studies of sine-wave speech synthesis. Remez, Rubin, Pisoni, and Carrell (1981) have shown that much, if not all, of the information for the perception of a novel utterance is preserved if the acoustic pattern, stripped of variations in overall amplitude and in the relative energy of formants, is reduced to a pattern of modulated sine waves following the approximate center frequencies of the three lowest formants. Here, it seems, nothing of the original signal is preserved other than changes, and derivatives of changes, in the frequency positions of the main peaks of the vocal tract transfer function (cf. Kuhn, 1975).

Finally, several recent audio-visual studies have shown that phonetic judgments of a spoken syllable can be modified if the listener simultaneously watches a video presentation of a face mouthing a different syllable: for example, a face uttering [ga] on video, while a loudspeaker presents [ba], is usually judged to be saying [da] (McGurk & MacDonald, 1976; Summerfield, 1979). The phonetic percept, in such a case, evidently derives from some combination of abstract, dynamic properties that characterize both auditory and visual patterns.

Moreover, infants are sensitive to dynamic correspondences between speech heard and speech seen. Three-month-old infants look longer at the face of a woman reading nursery rhymes if auditory and visual displays are synchronized, than if the auditory pattern is delayed by 400 milliseconds (Dodd, 1979). This finding evidently reflects more than a general preference for audiovisual synchrony, since six-month-old infants also look longer at the video display of a face repeating a disyllable that they hear (e.g., [lulu]) than at the synchronized display of a face repeating a different disyllable (e.g., [mama]) (MacKain, Studdert-Kennedy, Spieker, & Stern, Note 1).

The point here is not the cross-modal transfer of a pattern, which can be demonstrated readily in lower animals. Rather, it is the inference from this cross-modal transfer, and from the other evidence cited, that the speech signal conveys information about articulation by means of an abstract (and therefore modality-free) dynamic pattern. The infant studies hint further that the infant learns to speak by discovering its capacity to transpose that pattern into an organizing scheme for control of its own vocal apparatus.

Here we should note that, while the capacity to imitate general motor behavior may be quite common across animal species, a capacity for vocal imitation is rare. We should also distinguish social facilitation and general observational learning from the detailed processes of imitation, evidenced by the cultural phenomenon of dialects among whales, seals, certain songbirds, and humans. Finally, we should note that speech (like musical performance and, perhaps, dance) has the peculiarity of being organized, at one level of execution, in terms of a relatively small number of recurrent and, within limits, interchangeable gestures. Salient among these gestures are those that correspond to the processes of closing and opening the vocal tract, that is, to the onsets (or offsets) and to the nuclei of syllables.

We do not have to suppose that the child must analyze adult speech into features, segments, syllables, or even words, before she can set about imitating what she has heard. To suppose this would be to posit for speech a mode of development that precisely reverses the normal (phylogenetic and ontogenetic) process of differentiation. And, in fact, the earliest utterances used for symbolic or communicative ends seem to be prosodic patterns, which retain their unity across a wide variety of segmental realizations (Menn, 1976). Moreover, the early words also seem to be indivisible: for example, the child commonly pronounces certain sounds correctly in some words, but not in others (Menyuk & Menn, 1979). This implies that the child's first pass at the adult model of a word is an unsegmented sweep, a rough, analog copy of the unsegmented syllable. And there is no reason to believe that the child's percept is very much more differentiated than her production. Differentiation begins perhaps, when, with the growth of vocabulary, recurrent patterns emerge in the child's motor repertoire. Words intersect, and similar control patterns coalesce into more or less invariant segments. The segmental organization is then revealed to the listener by the child's distortions. Menn (1978, 1980) describes these distortions as the result of systematic constraints on the child's output: the execution of one segment of a word is distorted as a function of the properties of another. She classifies these constraints in terms of consonant harmony (e.g., [gʌk] for duck), consonant sequence (e.g., [nos] for snow), relative position (e.g., [dæge] for 'gator), and absolute position (e.g., [ɪʃ] for fish).

Here we touch on deep issues concerning the origin and nature of phonological rules. But the descriptive insights of Menn and others working in child phonology are important to the present argument because they seem to justify a view of the phonetic segment as emerging from recurrent motor patterns in the execution of syllables rather than as imposed by a specialized perceptual device. As motor differentiation proceeds, these recurrent patterns form classes, defined by their shared motor components--shared, in part, because the vocal tract has relatively few independently movable parts. These components are, of course, the motor origins of phonetic features (cf. Studdert-Kennedy & Lane, 1980). Some such formulation is necessary to resolve the paradox of a quasi-continuous signal carrying a segmented linguistic message. The signal carries no message: it carries information concerning its source. The message lies in the peculiar relation between the source and the listener, as a human and as a speaker of a particular language.

Readers familiar with the work of Turvey and Shaw (e.g., 1979) will recognize that the present sketch of a new approach to speech perception owes much to their ecological perspective (as also to Fowler, Rubin, Remez, &

Turvey, 1980). What may not be generally realized is that this perspective is highly compatible with much recent work in natural phonology (e.g., Stampe, 1979), child phonology (e.g., Menn, 1980), and phonetic theory (e.g., Lindblom, 1980; MacNeilage & Ladefoged, 1976; Ohala, in press). For example, Lindblom and his colleagues have, for several years, been developing principles by which the feature structure of the sound systems of different languages might be derived from perceptual and articulatory constraints. More generally, Lindblom (1980) has stressed that explanatory theory must refer "...to principles that are *independent* of the domain of the observations themselves" (p. 18) and has urged that phonetic theory "...move [its] search for basic explanatory principles into the physics and physiology of the brain, nervous system and speech organs..." (p. 18). In short, if language is a window on the mind, speech is the thin end of an experimental wedge that will pry the window open. The next ten years may finally see the first steps toward a genuine biology of language.

REFERENCE NOTE

1. MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. *Cross-modal coordination in infants' perception of speech*. Paper presented at the International Conference on Child Psychology, Vancouver, B.C., August 1981.

REFERENCES

Darwin, C. J. The perception of speech. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception, Vol. VII: Language and speech*. New York: Academic Press, 1976, 175-226.

Dodd, B. Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, 1979, *11*, 478-484.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, 1980, *27*, 343-350.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production*. New York: Academic Press, 1980.

Kuhn, G. M. On the front cavity resonance and its possible role in speech perception. *Journal of the Acoustical Society of America*, 1975, *58*, 428-433.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Liberman, A. M., & Pisoni, D. B. Evidence for a special speech-perceiving subsystem in the human. In T. H. Bullock (Ed.), *Recognition of complex acoustic signals*. Berlin: Dahlem Konferenzen, 1977, 59-76.

Lindblom, B. The goal of phonetics, its unification and application. *Phonetica*, 1980, *37*, 7-26.

MacNeilage, P., & Ladefoged, P. The production of speech and language. In E. Carterette & M. P. Friedman (Eds.), *Handbook of perception* (Vol. VII): *Language and speech*. New York: Academic Press, 1976, 75-120.

McGurk, H., & MacDonald, J. Hearing lips and seeing voices. *Nature*, 1976, *264*, 746-748.

Menn, L. Pattern, control and contrast in beginning speech: A case study in the development of word form and function. Bloomington, Ind.: Indiana University Linguistics Club, 1976.

Menn, L. Phonological units in beginning speech. In A. Bell & J. B. Hooper (Eds.), Syllables and segments. Amsterdam, North-Holland, 1978.

Menn, L. Phonological theory and child phonology. In G. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), Child phonology perception and production (Vol. 1). New York: Academic Press, 1980, 23-41.

Menyuk, P., & Menn, L. Early strategies for the perception and production of words and sounds. In P. Fletcher & M. Garman (Eds.), Language acquisition. New York: Cambridge University Press, 1979, 49-70.

Miller, G. A. Spontaneous apprentices. New York: Seabury Press, 1977, Ch. 7.

Ohala, J. The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (Ed.), Speech production. New York: Springer-Verlag, in press.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. Speech perception without traditional speech cues. Science, 1981, 212, 947-950.

Repp, B. H. Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Haskins Laboratories Status Report on Speech Research, 1981, SR-67/68, this volume.

Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. Perceptual integration of acoustic cues for stop, fricative, and affricate manner. Journal of Experimental Psychology: Human Perception and Performance, 1978, 4, 621-637.

Stampe, D. A dissertation on natural phonology. New York: Garland, 1979.

Studdert-Kennedy, M. Speech perception. In N. J. Lass (Ed.), Contemporary issues in experimental phonetics. New York: Academic Press, 1976, 243-293.

Studdert-Kennedy, M. Speech perception. Language and Speech, 1980, 23, 45-66.

Studdert-Kennedy, M., & Lane, H. The structuring of language: Clues from the differences between signed and spoken language. In U. Bellugi & M. Studdert-Kennedy (Eds.), Signed language and spoken language: Biological constraints on linguistic form. Deerfield Beach, Fl.: Verlag Chemie, 1980, 29-40.

Summerfield, Q. Use of visual information for phonetic perception. Phonetica, 1979, 36, 314-331.

Summerfield, Q., & Haggard, M. On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. Journal of the Acoustical Society of America, 1977, 62, 436-448.

Turvey, M. T., & Shaw, R. E. The primacy of perceiving: An ecological reformulation of perception for understanding memory. In L-G. Nilsson (Ed.), Perspectives on memory research: Essays in honor of Uppsala University's 500th anniversary. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1979.

AUDITORY INFORMATION FOR BREAKING AND BOUNCING EVENTS: A CASE STUDY IN ECOLOGICAL ACOUSTICS

William H. Warren, Jr.+ and Robert R. Verbrugge+

Abstract. The mechanical events of bouncing and breaking glass are acoustically specified by single vs. multiple damped quasi-periodic pulse patterns, with an initial noise burst in the case of breaking. Subjects show high accuracy in distinguishing natural tokens of these two events and tokens constructed by adjusting the periodicities of spectrally identical components. Differences in average spectral frequency are therefore not necessary for perceiving this contrast, though differences in spectral consistency over successive pulses apparently are important. Initial noise corresponding to glass rupture is not necessary to distinguish breaking from bouncing, but may be important for identifying breaking in isolation. The data indicate that higher-order temporal invariants in the acoustic signal provide information for the auditory perception of these events.

Research in auditory perception has emphasized the detection and processing of sound elements with quasi-stable spectral structure, such as tones, formants, and bursts of noise. In the spectral domain, these elements are distinguished by frequency peak or range, bandwidth, and amplitude. In the temporal domain, acoustic analysis has often been limited to the durations of sound elements and the intervals between them. Much of traditional perceptual research, including that of classical psychoacoustics, has focused on listeners' response accuracy to essentially time-constant functions of frequency, amplitude, and duration, on the assumption that complex auditory percepts are compositions over sound elements with those properties (Fletcher, 1934; Helmholtz, 1863/1954; Plomp, 1964; see Green, 1976).

The perceptual role of time-varying properties of sound has received comparatively little attention. Some exceptions to this can be found in research on amplitude and frequency modulation, particularly as they relate to classical auditory phenomena such as beats and periodicity pitch. In general, however, research on time-varying properties has been most common in the study

+Also University of Connecticut.

Acknowledgment. The authors wish to thank Michael Studdert-Kennedy, Michael Turvey, and Robert Shaw for their comments. This research was supported by a grant from the National Institute for Child Health and Human Development (HD-01994) and a Biomedical Research Support Grant (RR-05596) to Haskins Laboratories, and by a National Science Foundation graduate fellowship to the first author.

[HASKINS LABORATORIES: Status Report on Speech Research SR-67/68 (1981)]

of classes of natural events, such as human speech, music, and animal communication, where an analysis of sound into quasi-stable elements is often problematic. In the case of speech, for example, many phonemic contrasts can be defined by differences in the direction and rate-of-change of major speech resonances (see Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Delattre, Gerstman, & Cooper, 1956). Some research on the perception of music has also demonstrated the perceptual significance of time-varying properties. Identification of musical instruments, for example, is strongly influenced by the temporal structure of transients that accompany tone onsets (Luce & Clark, 1967; Saldanha & Corso, 1964). In particular, the relative onset timing and the rates of amplitude change of upper harmonics have been found to be critical properties of attack transients that permit distinctions among instrument families (Grey, 1977; Grey & Gordon, 1978). Animal vocalizations are similarly rich in time-varying properties (such as rhythmic pulsing, frequency modulation, and amplitude modulation), and many of these properties have been shown to be critical for distinguishing the species, sex, dangerousness, location, and motivational state of the producer (e.g., Brown, Beecher, Moody, & Stebbins, 1978; Konishi, 1978; Peterson, Beecher, Zoloth, Moody, & Stebbins, 1978).

It is noteworthy that in each of these areas of research on natural events, the discovery or explanation of perceptually significant, time-varying, acoustic properties has been motivated by an analysis of the time-varying behavior of the sound source. In the case of speech, for example, an analysis of speech production has been an integral part of the search for the acoustic basis for speech perception (e.g., Fant, 1960; Fowler, 1977; Fowler, Rubin, Remez, & Turvey, 1980; Liberman et al., 1967; Verbrugge, Rakerd, Fitch, Tuller, & Fowler, in press). It is also worth noting that researchers in these areas have often found it more useful to characterize perceptual information in terms of higher-order structure in sound--that is, in terms of functions over the traditional measures of frequency, amplitude, and duration. Given the time-varying behavior of the sound sources involved, it is not surprising that many of these functions are time-dependent in nature, defining rates of change and styles of change in lower-order acoustic variables. Finally, it is not uncommon for researchers in these fields to view this temporal structure as a property of the sound stream itself, rather than as a property that must be introduced by a perceiver while constructing a percept.

The role of time-varying properties in the perception of other familiar events in the human environment is largely unknown, and research on the subject has been sparse. Our goal in this paper is to demonstrate by argument and example that higher-order, temporal structure can be important for distinguishing such events.

It is apparent from everyday experience that listeners can detect significant aspects of the environment by ear, from a knock at the door to the condition of an automobile engine and the gait of an approaching friend. Such naturalistic observations were recently verified in experiments by VanDerveer (Note 1, Note 2). She presented 30 recorded items of natural sound in a free identification task and found that many events such as clapping, footsteps, jingling keys, and tearing paper were identified with greater than 95% accuracy. Subjects tended to respond by naming a mechanical event that produced the sound, and reported their experiences in terms of sensory qualities only when source recognition was not possible. VanDerveer (Note 1) also found that confusion errors in identification tasks and clustering in sorting tasks tended to group acoustic events by common temporal patterns. For example, hammering was confused with walking, and the scratching of fingernails was confused with filing, but hammering and walking were not confused with the latter two events.

These results support the general claim that sound in isolation permits accurate identification of classes of sound-producing events when the temporal structure of the sound is specific to the mechanical activity of the source (Gibson, 1966b; Schubert, 1974; Warren & Verbrugge, in press). If higher-order information is found to be specific to events, while values of lower-order variables per se are not, then it may be more fruitful to view the auditory system as being designed for the perception of source events (via higher-order acoustic functions), rather than for the detection of quasi-stable sound elements. Schubert (1974) put this succinctly in his "Source Identification Principle" for auditory perception: "Identification of sound sources, and the behavior of those sources, is the primary task of the [auditory] system" (p. 126).

This general perspective on auditory perception is coming to be called "ecological acoustics," on a direct analogy to the ecological optics advocated by Gibson (1961, 1966b) as an approach to vision. The ecological approach leads to research that is similar in many respects to the work summarized above on speech, music, and animal communication. In general terms, the strategy for research is to identify the higher-order properties that are defined over the course of a natural sound-producing event, and then to assess the ability of listeners to utilize that potential information. A physical analysis of the source and its behavior is an essential part of the strategy, both for identifying acoustic variables that might otherwise be missed, and for bounding the set of possible variables in a principled fashion. Furthermore, demonstrating the specificity of acoustic structure to the source event is crucial to avoid the introduction of ad hoc processing principles to buttress perception (Shaw, Turvey, & Mace, in press).

In addition to offering a research strategy, the ecological approach seeks a general analysis of events and a description of the perceptual information specific to them. This analysis is based on the observation that identifiable objects participate in identifiable transformations or "styles of change" (Gibson, 1966a; Pittenger & Shaw, 1975; Shaw & Cutting, 1980; Shaw, McIntyre, & Mace, 1974; Shaw & Pittenger, 1978; Johansson, Hofsten, & Jansson, 1980). More precisely, a class of objects may be functionally defined in terms of structure that is preserved and destroyed under certain transformations. The information that specifies the kind of object and its properties under change is known as the structural invariant of an event (Pittenger & Shaw, 1975). Reciprocally, the information that specifies the style of change is known as the transformational invariant, which may be described jointly in terms of the geometric properties that remain constant and those that vary systematically under change (Pittenger & Shaw, 1975; Mark, Todd, & Shaw, in press).

By such an analysis, events can be organized into equivalence classes ("types") that are defined by sets of transformational and structural invari-

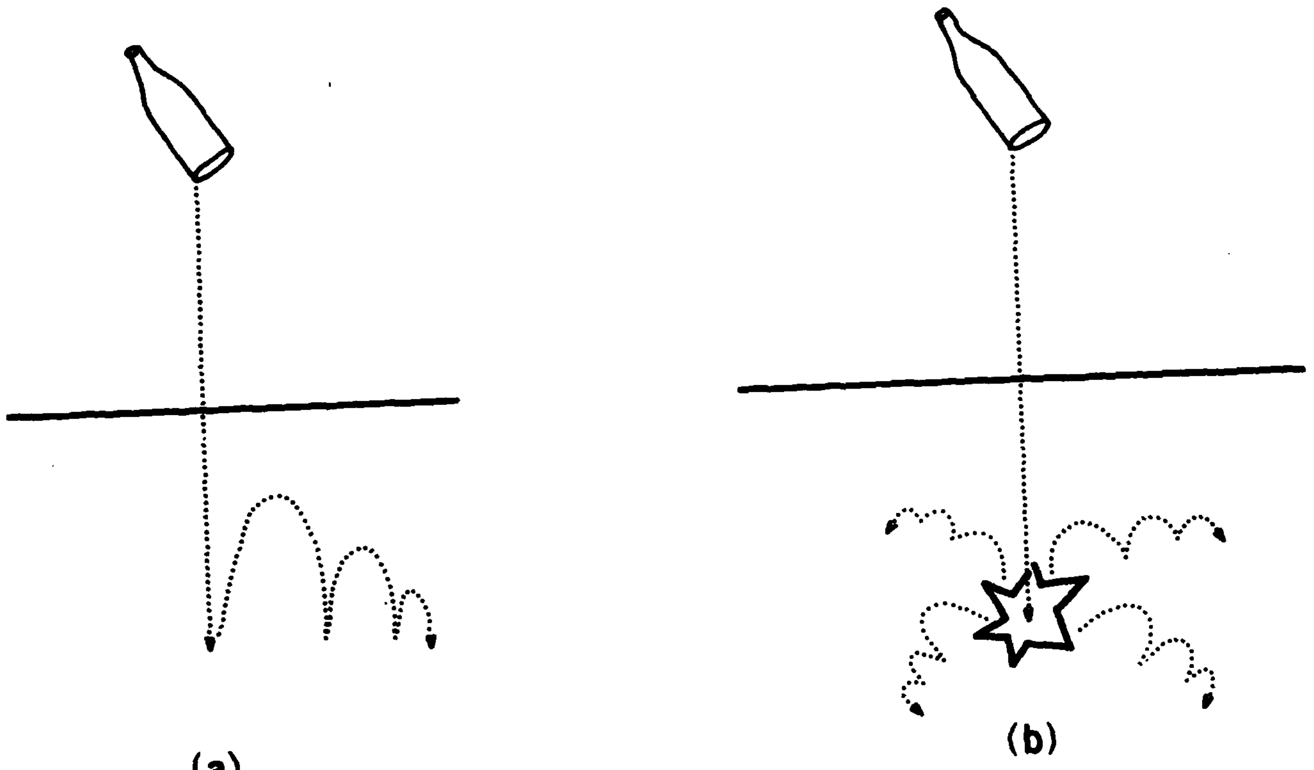


Figure 1. Cartoons of the mechanical events of (a) bouncing, and (b) breaking.

ants. Consider, for example, the style of change of walking and the animals with appropriate limb structures, or the style of change of burning and the objects that are combustible under terrestrial conditions. Within any equivalence class of events, an indefinite number of particular instances ("tokens") are possible, each preserving the invariants of the class but individuated in space and time--that charging rhino, or this burning bridge. For any perceptible event, information about its class membership is, by hypothesis, available by means of the physical media it disturbs. An analysis of such potential information and its relationship to the source event is a major goal of ecological physics.

The present paper explores the acoustic aspects of dropping a glass object and its subsequent bouncing or breaking. Bouncing and breaking are two distinct styles of change that may be wrought over a variety of objects, such as bottles, plates, pottery, and other ceramics. These two events would be identified by Gibson (1979, pp. 94-95) as changes of the layout of surfaces due to physical force--bouncing as a case of successive collisions, breaking as a compound event of surface rupturing followed by successive collisions (and possible further rupturing) of the broken pieces. The two styles of change constitute disjoint equivalence classes of events: the breaking and the bouncing of semi-elastic objects. By acoustic and perceptual studies of these events, we hope to discover the transformational invariants that distinguish them. (Structural invariants specifying individual properties of the objects such as size, shape, and material, and individual transformation properties such as height of drop, force of impact, and angle of impact, are discussed in Warren & Verbrugge, in press.)

Consider first the mechanical action of a bottle bouncing on a hard surface (see Figure 1a). Each collision consists of an initial impact that briefly sets the bottle into vibration at a set of frequencies determined by its size, shape, and material composition. This is reflected in the acoustic signal as an initial burst of noise followed by spectral energy concentrated at a particular set of overtone frequencies. Over a series of bounces, the collisions between object and ground occur with declining impact force and decreasing ("damped") period, although some irregularities in the pattern may occur due to the asymmetry of the bottle. The spectral components are similar across bounces, relative overtone amplitudes varying slightly due to the varying orientations of the bottle at impact. (The spectrum within each pulse is quasi-stable, and is conventionally described in terms of spectral peaks in a cross-section of the signal.) These acoustic consequences may be described as a single damped quasi-periodic pulse train in which the pulses share a similar cross-sectional spectrum (Figure 2a). It is this single pulse train that we suggest constitutes a transformational invariant of temporal patterning for the bouncing style of change.

Turning to the mechanical action of breaking (Figure 1b), it is evident that a catastrophic rupture occurs upon impact. Assuming an idealized case, the resulting pieces then continue to bounce without further breakage, each with its own independent collision pattern. The acoustic consequences appear as an initial rupture burst dissolving into overlapping multiple damped quasi-periodic pulse trains, each train having a different cross-sectional spectrum and damping characteristic (Figure 2b). We propose that a compound signal, consisting of a noise burst followed by such multiple pulse trains, consti-

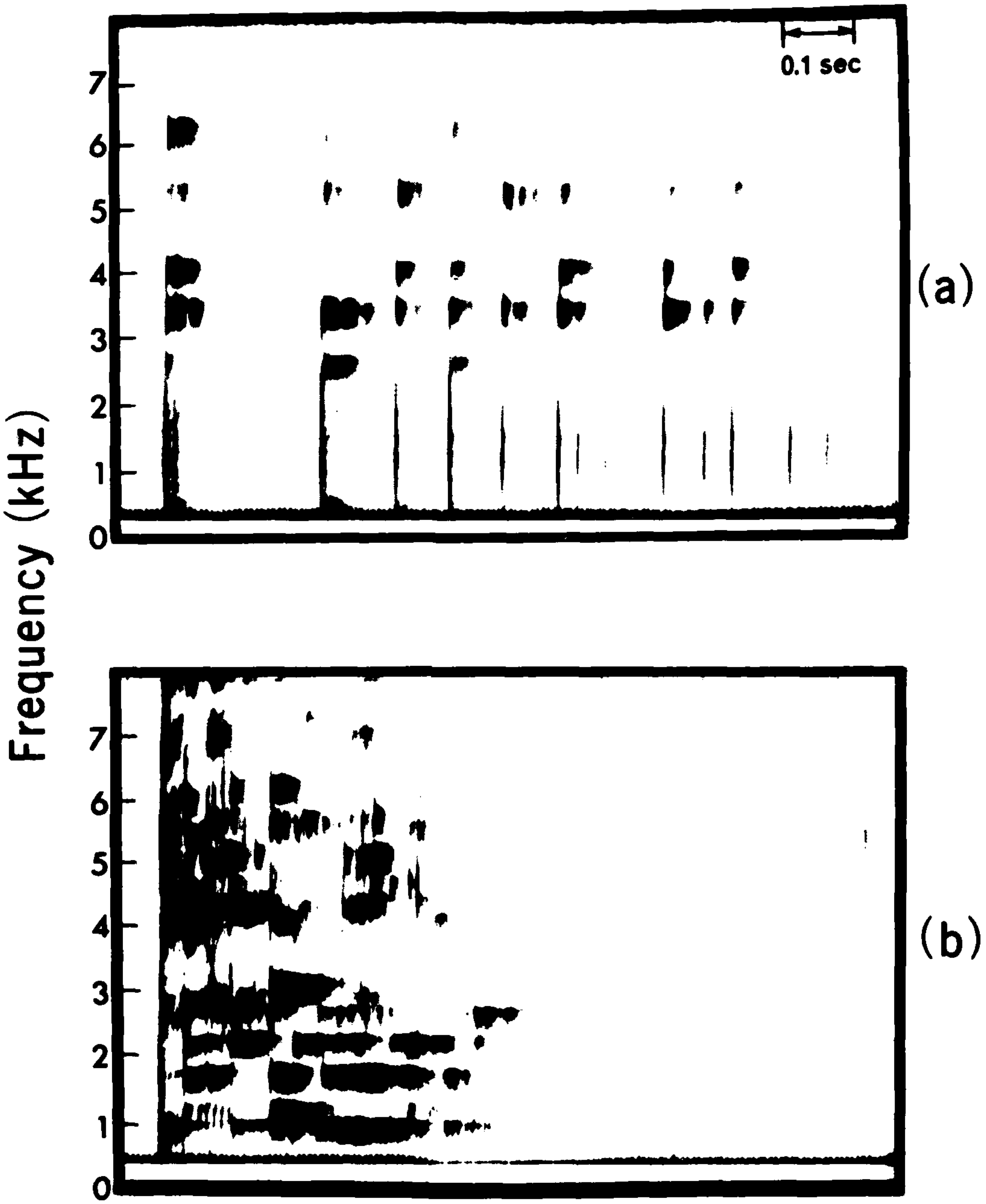


Figure 2. Spectrograms of natural tokens: (a) bouncing, (b) breaking.

tutes the transformational invariant that specifies the compound event of breaking.

Aside from these aspects of temporal pattern and initial noise, certain crude spectral differences between breaking and bouncing can be observed by comparing spectrograms of natural cases (Figure 2). First, the overtones of breaking events are distributed across a wider range of frequencies than are those of bouncing events. Second, the overtones of breaking are denser in the frequency domain. Both of these properties can be traced to the contrast between a single object in vibration and a number of disparate objects simultaneously in vibration.

The following experiments test the hypothesis that temporal patterning, rather than some quasi-stable spectral property, distinguishes breaking from bouncing. By superimposing recordings of individual pieces of broken glass, cases of breaking and bouncing can be constructed from a common set of pieces by varying the temporal correspondences among their collision patterns. Experiment 1 establishes that listeners can identify natural cases of breaking and bouncing with high accuracy. Experiment 2 examines performance on constructed cases that include an initial breakage burst, and compares it with the results for natural sound. Finally, Experiment 3 assesses the contribution of the burst by removing it from both natural and constructed cases of breaking.

EXPERIMENT 1: NATURAL SOUND

The first experiment determines whether natural sound provides sufficient acoustic information for listeners to distinguish the events of breaking and bouncing.

Method

Materials. Natural recordings were made of three glass objects dropping onto a concrete floor covered by linoleum tile in a sound-attenuated room. Using a Crown 800 tape deck, the sound of each object was recorded when the object was dropped from a 1 ft. height (bouncing), and when it was dropped from a 2 to 5 ft. height (breaking). This yielded three tokens of bouncing and three tokens of breaking. The objects used and the durations of the bouncing (BNC) and breaking (BRK) events are as follows: (1) 32 oz. jar: BNC1 = 1600 msec, 22 bounces; BRK1 = 1200 msec. (2) 64 oz. bottle: BNC2 = 1600 msec, 15 bounces; BRK2 = 550 msec. (3) 1 litre bottle: BNC3 = 1300 msec, 17 bounces; BRK3 = 700 msec. The recordings were digitized at a 20 kHz sampling rate using the Pulse Code Modulation (PCM) system at Haskins Laboratories. A test tape was then recorded; it contained 20 trials of each natural token in randomized order for a total of 120 test trials. A pause of 3 sec occurred between trials, and a pause of 10 sec occurred after every six trials.

Subjects. Fifteen graduate and undergraduate students participated in the experiment for payment or course credit.

Procedure. Subjects were run in groups of two to five and listened to the tape binaurally through headphones. They were told that they would be hearing recordings of objects that had either bounced or broken after being dropped, but were told nothing about the nature of the objects involved. Their three-choice task was to identify each event as a case of breaking or bouncing, with a "don't know" option, by placing a check in the appropriate column on an answer sheet. The "don't know" category was included to minimize the possibility that subjects would choose one of the two event categories even when they found the sound unconvincing, as they would be forced to do in a two-choice situation. They were specifically instructed to ignore the nature of the object involved and attend to "what's happening to it." Subjects received no practice trials or feedback. There was a short break after 60 trials, and a test session lasted about 20 min.

Results and Discussion

Overall performance on natural bouncing tokens was 99.3% correct ("bouncing" judgments), and on breaking tokens was 98.5% correct ("breaking" judgments). "Don't know" responses accounted for 0.2% of all answers on bouncing tokens and 0.7% on breaking tokens. Experiment 1 clearly demonstrates that sufficient information is present in the acoustic signal to permit unpracticed listeners to distinguish the events of bouncing and breaking.

EXPERIMENT 2: CONSTRUCTED SOUND

Experiment 2 attempted to model the time-varying information contained in natural recordings by using constructed cases of bouncing and breaking, eliminating average spectral differences between the two.

Method

Materials. Tokens intended to model bouncing and breaking were constructed by the following method. Initially, individual recordings were made of four major pieces of glass from a broken bottle as each piece was dropped and bounced separately from a low height. These recordings were combined in two ways using the PCM system.

To construct a bouncing token, the temporal pattern of each piece was adjusted to match a single master periodicity arbitrarily borrowed from a recording of a natural bouncing bottle (Figure 3a). This was accomplished by inserting tape hiss between the bounce pulses in recordings of the individual pieces. After all four pieces had been adjusted so that their onsets matched the same pulse pattern, they were superimposed by summing the instantaneous amplitudes of the digitized recordings. The result was a combined pulse pattern with synchronized onsets for all bounces, preserving the invariant of a single damped quasi-periodic pulse train to model bouncing (Figure 4a).

A breaking token was constructed by readjusting the same four pieces to match four different temporal patterns (Figure 3b). As a first approximation, these master patterns were borrowed from measurements of four different bouncing bottles, since the likely patterns of individual pieces of glass in the course of natural breaking were unknown. These four patterns were initiated simultaneously, preceded by 50 msec of noise burst taken from the original rupture. The result after superimposing these four independent

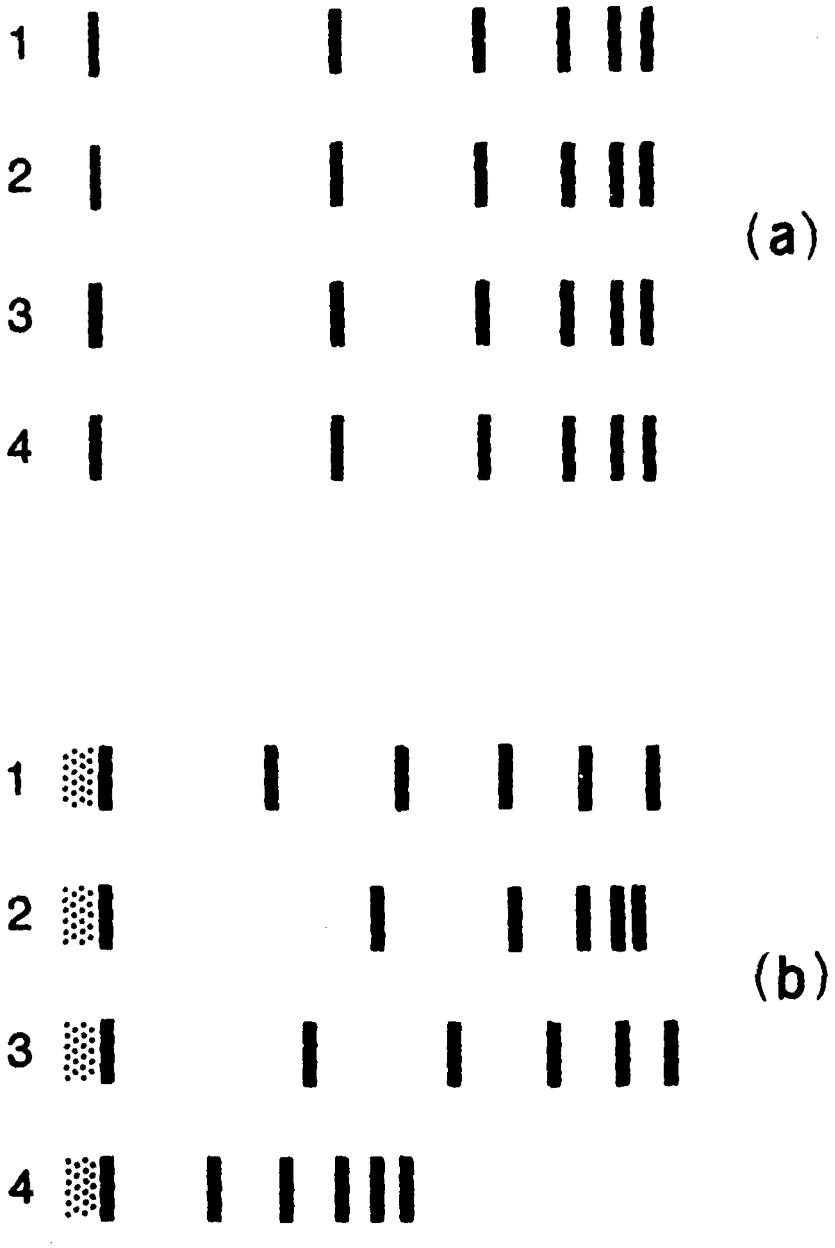


Figure 3. Schematic diagrams of constructed tokens, combining four component pulse trains: (a) bouncing, with synchronous pulse onsets, (b) breaking, with initial noise burst and asynchronous pulse onsets.

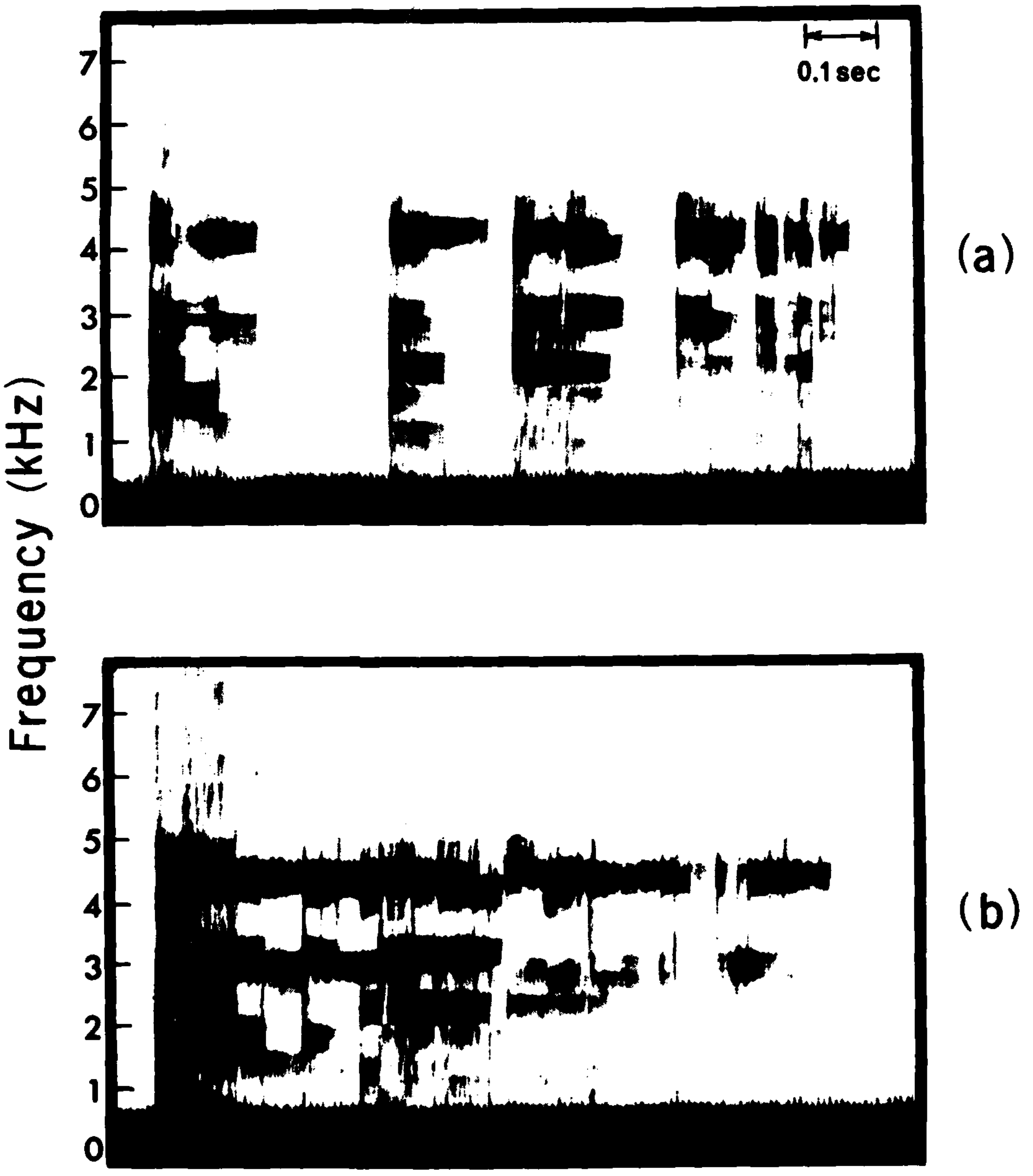


Figure 4. Spectrograms of constructed tokens: (a) bouncing (SYN1), (b) breaking (ASYN1).

temporal series was a combined pattern with asynchronous pulse onsets, preserving the temporal invariant of multiple damped quasi-periodic pulse trains to model breaking (Figure 4b). Note that the variables of temporal patterning and initial noise were confounded in this experiment. To the experimenters' ears the burst improved the quality of apparent breakage, but this assumption was later tested in Experiment 3.

Hence, the only differences between bouncing and breaking tokens were in the temporal registration of pulse onsets and the presence (or absence) of initial noise. The range and distribution of average spectral frequencies were similar in the two cases. Mean overall durations differed, averaging 1107 msec for bouncing tokens and 733 msec for breaking tokens; in general this factor is related to object elasticity and the height of drop, and it is therefore not a likely candidate for information specific to a style of change.

There were certain problems with the constructed cases. The process of superimposing pulse patterns also summed tape hiss and hum, so that background noise was increased. Moreover, constructing the sound of a single bouncing object by combining the spectral components of four independently bouncing pieces produced in one case a noise that sounded more like metal than glass material; nevertheless, the temporal invariant was preserved. The other two bouncing tokens sounded like glass. Finally, the use of only four pieces of glass to simulate breaking, the assumption that their periodicities were akin to those of a bouncing bottle, and the assumption of no further breakage after the initial catastrophe, were all rather arbitrary idealizations. Nevertheless, if temporal patterning constitutes information for breaking and bouncing, subjects should be able to make reliable judgments of these tokens.

Three cases of bouncing and three corresponding cases of breaking were produced by this method, each pair constructed from a unique set of original pieces and matched to a unique set of master periodicities. The original objects, and the durations of the bouncing or synchronous (SYN) and breaking or asynchronous (ASYN) tokens constructed from their pieces, were as follows: (1) 32 oz. jar: SYN1 = 1000 msec, 8 bounces; ASYN1 = 950 msec. (2) 32 oz. jar: SYN2 = 1400 msec, 13 bounces; ASYN2 = 650 mec. (3) 64 oz. bottle: SYN3 = 920 msec, 9 bounces; ASYN3 = 600 msec.

Subjects. Fifteen graduate and undergraduate students participated in the experiment for payment or course credit. None of them had participated in Experiment 1.

Procedure. The procedure was the same as that in Experiment 1, with the exception that trials were presented in blocks of ten rather than blocks of six. Instructions to the subjects were the same, including the instruction to ignore object properties and concentrate on the style of change.

Results and Discussion

The results for each constructed token appear in Table 1, and are consistent with the predictions of the temporal patterning hypothesis. Bouncing judgments on synchronous tokens averaged 90.7%, and breaking judgments on asynchronous tokens averaged 86.7% (these judgments being treated as

Table 1

Percent Correct Judgments on Constructed Tokens

(Experiment 2)

Token	Bouncing	Breaking
1	93.7	89.0
	(18.7, 1.7)	(17.8, 2.9)
2	94.3	71.3
	(18.9, 2.0)	(14.3, 4.4)
3	84.3	99.7
	(16.9, 3.7)	(19.9, 0.3)
Overall	90.7	86.7

Note: Mean scores and standard deviations are in parentheses. Scores are based on 20 trials per subject per cell, N=15.

"correct"). "Don't know" answers accounted for 0.1% of all responses on bouncing tokens, and 1.3% on breaking tokens. Considering the artificial nature of the constructed cases and the idealizations involved, their identifiability may be considered quite high.

Some departures from the general pattern were found for token ASYN2, which showed a markedly lower level of correct performance (71.3%), a higher standard deviation for "breaking" judgments, and a relatively high rate of "don't know" responses (4.0%). These differences were primarily due to the low performance of five subjects who averaged 44% correct on this token, while the performance of the other ten averaged 85.0%. It may be noted that the summed background noise was greater in ASYN2 than in the other two breaking cases. The fact that overall performance in this case was well above chance indicates that even the token of lowest identifiability contained sufficient information to distinguish the two events.

It is not surprising that some tokens of constructed breaking are more convincing than others, as there are certainly some natural instances that are more compelling than others. The differences among tokens may involve both the spectral distinctiveness of the broken pieces and their degree of asynchrony. In pilot tests, when the pulses from a single piece were adjusted to four different periodicities, the resulting sum of the four patterns did not specify breaking. Apparently, distinct spectral properties for each piece are necessary to distinguish multiple pulse trains (see Figure 1b). The reciprocal bouncing case, in which successive pulses were borrowed from different bottles, similarly failed to yield a coherent bouncing event. Hence, spectral similarity across pulses appears to be necessary to specify the unity of a single pulse train.

In general, performance with constructed sound was similar to that found for natural sound in Experiment 1. Although performance with constructed cases was somewhat lower than with natural cases, the differences were only about 10% on average, and performance with both natural and constructed cases was far above the chance level. The data permit us to conclude that temporal patterning is compelling information for breaking and bouncing. In other words, constructed and natural cases appear to specify the same general equivalence classes of breaking and bouncing events to a listener.

EXPERIMENT 3: INITIAL NOISE SPECIFIC TO RUPTURING

To isolate the variable of single vs. multiple pulses and assess the importance of the initial noise burst in specifying breaking, the first two experiments were repeated with initial noise removed from both natural and constructed cases. Pilot work indicated that the first 80 msec of the signal in natural breaking and bouncing cases was not, in isolation, sufficient to distinguish the two events. Experiment 3 was conducted to determine whether the initial noisr, in addition to the pulse patterns, was necessary to distinguish breaking from bouncing.

Method

Materials. Both natural and constructed tokens were prepared. Bouncing tokens were the same as those used in the two previous experiments. For breaking cases, the constructed tokens from Experiment 2 were modified by removing the 50 msec of initial noise that had been added for that experiment. The natural breaking tokens from Experiment 1 were modified by removing the naturally occurring burst. Since there was no distinct boundary in the natural waveform between the rupture burst and the subsequent collision pulses, the natural tokens were edited by removing noise identifiable on an oscillogram and by listening for the absence of a burst. This technique resulted in the removal of the initial 80 msec from BRK1, 50 msec from BRK2, and 60 msec from BRK3. In sum, there were three tokens of bouncing and three tokens of breaking (without initial noise) in both the natural and constructed conditions.

Subjects. Thirty graduate and undergraduate students participated in the experiment for payment. None had participated in the previous experiment.

Procedure. The natural and constructed conditions were run separately with two different groups of 15 subjects. The procedure and instructions were the same as before, with each group receiving 120 randomly ordered trials in blocks of six.

Results and Discussion

The results for each token appear in Table 2. With natural cases, the overall performance was 99.8% correct on bouncing tokens and 96.0% correct on breaking tokens; with the constructed cases it was 93.0% for bouncing and 86.7% for breaking. These results were nearly identical to those of Experiment 1 with natural sound and Experiment 2 with constructed sound. "Don't know" answers accounted for 0.0% of all responses on natural bouncing tokens, 1.0% on natural breaking tokens, 2.0% on constructed bouncing tokens, and 4.0% on constructed breaking tokens.

Hence, removal of initial noise from breaking tokens does not reduce their discriminability. Finding this result for the natural cases indicates that the burst is not necessary to distinguish the two events. The same finding with constructed cases demonstrates that variation in the temporal patterning of pulse onsets is alone sufficient to discriminate breaking and bouncing.

However, we may question whether pulse patterning alone is sufficient to specify a breaking event in isolation. Following the test sessions, a number of subjects in Experiment 3 reported that natural and constructed breaking cases without initial noise often provided weak instances of the event, some sounding more like "wind chimes," "bells," "spoons dropping," or "ice cubes in a glass"—in other words, like multiple collisions of initially independent objects. Others reported precisely what was presented: "pieces falling after the break, without an initial crash." Although the acoustic structure was sufficient to distinguish the event of breaking from that of bouncing, and not ambiguous enough to merit a "don't know," it could nevertheless specify wind chimes, not breaking glass, when heard in isolation. Since breaking is a

Table 2

Percent Correct Judgments on Natural and Constructed Tokens

Without Initial Noise (Experiment 3)

	Natural		Constructed	
Token	Bouncing	Breaking	Bouncing	Breaking
1	99.7	93.7	94.0	83.7
	(19.9, 0.3)	(18.7, 2.4)	(18.8, 2.2)	(16.7, 2.9)
2	99.7	97.7	97.3	76.7
	(19.9, 0.3)	(19.5, 1.6)	(19.5, 0.9)	(15.3, 4.1)
3	100.0	96.7	87.7	99.7
	(20.0, 0.0)	(19.3, 2.3)	(17.5, 2.7)	(19.9, 0.3)
Overall	99.8	96.0	93.0	86.7

Note: Mean scores and standard deviations are in parentheses. Scores are based on 20 trials per subject per cell, N=15 in the Natural condition and N=15 in the Constructed condition.

compound event, it is not surprising that the causal transition from one to many pieces must be specified by an initial rupture noise.

In general, these observations are consistent with our original hypothesis that breaking is specified by a complex acoustic configuration, consisting of an initial noise followed by multiple quasi-periodic pulse trains. Further work remains to be done to determine whether the initial noise is necessary for identifying breakage under conditions less constrained than in the present experiment.

GENERAL DISCUSSION

The preceding experiments have attempted to determine whether higher-order, time-varying properties constitute effective acoustic information for the events of bouncing and breaking. The results show that differences in the temporal patterning of component pulse onsets are sufficient to perceptually distinguish the two events, with or without an initial burst. These temporal invariants override any contribution of average spectral properties in distinguishing the events. The results provide evidence that certain damped periodic patterns, plus initial noise, constitute transformational invariants that specify breaking and bouncing to a listener.

However, if these temporal patterns are to convey the distinct events of breaking and bouncing, they must be carried by signals with certain spectral properties. Specifically, a single damped quasi-periodic pulse train must be of constant resonance if it is to cohere as the bouncing of a single object. Reciprocally, multiple damped quasi-periodic pulse trains must have different frequency spectra if they are to separate perceptually as independently bouncing shards, which together specify the breaking of an object into pieces. Hence, a combination of temporal and spectral patterns constitutes the information necessary and sufficient to specify breaking and bouncing.

The amplitude and periodicity requirements of such patterns in bouncing events were considered in two simple demonstrations worth mentioning here. Iterating a recording of one bounce pulse to match the timing of a natural bouncing sequence produced a clear bouncing event, although the usual declining amplitude gradient was absent. However, adjusting the pulse pattern to create equal 100 msec intervals between pulse onsets, thereby eliminating the damping of the periodic pattern, destroyed the effect of perceived bouncing. The rapid staccato sound was like that produced by a negentropic machine, such as a jackhammer. A damped series of collisions, as constrained by gravity and the imperfect elasticity of the system, appears necessary to the information for bouncing. Experiments are in progress to assess the efficacy of period damping in specifying elasticity or "bounciness" itself.

The experiments exemplify an ecological approach to auditory perception, seeking to identify higher-order acoustic information for complex events. The acoustic consequences of two distinct mechanical events were analyzed for their temporal and spectral structure, and the invariant properties sufficient to convey aspects of the events to a listener were empirically determined. Such work is preliminary to modeling auditory mechanisms capable of detecting these invariants (see Mace, 1977).

REFERENCE NOTES

1. VanDerveer, N. J. Acoustic information for event perception. Paper presented at the Celebration in Honor of Eleanor J. Gibson, Cornell University, June 1979.
2. VanDerveer, N. J. Confusion errors in identification of environmental sounds. Paper presented at the meeting of the Acoustical Society of America, Cambridge, Massachusetts, June 1979.

REFERENCES

Brown, C. H., Beecher, M. D., Moody, D. B., & Stebbins, W. C. Localization of primate calls by Old World monkeys. Science, 1978, 201, 753-754.

Fant, G. Acoustic theory of speech production. The Hague: Mouton, 1960.

Fletcher, H. Loudness, pitch, and the timbre of musical tones and their relation to the intensity, the frequency, and the overtone structure. Journal of the Acoustical Society of America, 1934, 6, 59-69.

Fowler, C. A. Timing control in speech production. Unpublished doctoral dissertation, University of Connecticut, Storrs, 1977.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. Implications for speech production of a general theory of action. In B. Butterworth (Ed.), Language production. New York: Academic Press, 1980.

Gibson, J. J. Ecological optics. Vision Research, 1961, 1, 253-262.

Gibson, J. J. The problem of temporal order in stimulation and perception. Journal of Psychology, 1966, 62, 141-149. (a)

Gibson, J. J. The senses considered as perceptual systems. Boston: Houghton Mifflin, 1966. (b)

Gibson, J. J. The ecological approach to visual perception. Boston: Houghton Mifflin, 1979.

Green, D. M. An introduction to hearing. Hillsdale, NJ: Erlbaum, 1976.

Grey, J. M. Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 1977, 61, 1270-1277.

Grey, J. M., & Gordon, J. W. Perceptual effects of spectral modifications on musical timbres. Journal of the Acoustical Society of America, 1978, 63, 1493-1500.

Johansson, G., Hofsten, C. von, & Jansson, G. Event perception. Annual Review of Psychology, 1980, 31, 27-63.

Helmholtz, H. L. F. von On the sensations of tone as a physiological basis for the theory of music (A. J. Ellis, trans.). New York: Dover, 1954. (Originally published, 1863).

Konishi, M. Ethological aspects of auditory pattern recognition. In R. Held, H. Leibowitz, & H. L. Teuber (Eds.), Handbook of sensory physiology, v. VIII: Perception. New York: Springer-Verlag, 1978.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. Psycholog cal Review, 1967, 74, 431-461.

Liberman, A. M., Delattre, P. C., Gerstman, L. T., & Cooper, F. S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. Journal of Experimental Psychology, 1956, 52, 127-137.

Luce, D., & Clark, M. Physical correlates of brass-instrument tones. Journal of the Acoustical Society of America, 1967, 42, 1232-1243.

Mace, W. M. James Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In R. E. Shaw & J. Bransford

(Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology*. Hillsdale, NJ: Erlbaum, 1977.

Mark, L. S., Todd, J. T., & Shaw, R. E. The perception of growth: How different styles of change are distinguished. *Journal of Experimental Psychology: Human Perception and Performance*, in press.

Petersen, M. R., Beecher, M. D., Zoloth, S. R., Moody, D. E., & Stebbins, W. C. Neural lateralization of species-specific vocalizations by Japanese macaques (*Macaca fuscata*). *Science*, 1978, *202*, 324-327.

Pittenger, J. B., & Shaw, R. E. Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, *1*, 374-382.

Plomp, R. The ear as a frequency analyzer. *Journal of the Acoustical Society of America*, 1964, *36*, 1628-1636.

Saldanha, E. L., & Corso, J. F. Timbre cues and the identification of musical instruments. *Journal of the Acoustical Society of America*, 1964, *36*, 2021-2026.

Schubert, E. D. The role of auditory perception in language processing. In D. D. Duane & M. B. Rawson (Eds.), *Reading, perception, and language*. Baltimore: York Press, 1974.

Shaw, R. E., & Cutting, J. E. Cues from an ecological theory of event perception. In U. Bellugi & M. Studdert-Kennedy (Eds.), *Signed language and spoken language: Biological constraints on linguistic form*. Weinheim: Verlag Chemie, 1980.

Shaw, R. E., & Pittenger, J. B. Perceiving change. In H. L. Pick & E. Saltzman (Eds.), *Modes of perceiving and processing information*. Hillsdale, NJ: Erlbaum, 1978.

Shaw, R. E., McIntyre, M., & Mace, W. M. The role of symmetry in event perception. In R. B. MacLeod & H. L. Pick (Eds.), *Perception: Essays in honor of James J. Gibson*. Ithaca, NY: Cornell University Press, 1974.

Shaw, R. E., Turvey, M. T., & Mace, W. M. Ecological psychology: The consequence of a commitment to realism. In W. Weimer & D. Palermo (Eds.), *Cognition and the symbolic processes, II*. Hillsdale, NJ: Erlbaum, in press.

Verbrugge, R. R., Rakerd, B., Fitch, H., Tuller, B., & Fowler, C. A. Perception of speech events: An ecological approach. In R. E. Shaw & W. M. Mace (Eds.), *Event perception: An ecological perspective*. Hillsdale, NJ: Erlbaum, in press.

Warren, W. H., & Verbrugge, R. R. Toward an ecological acoustics. In R. E. Shaw & W. M. Mace (Eds.), *Event perception: An ecological perspective*. Hillsdale, NJ: Erlbaum, in press.

SPEECH AND SIGN: SOME COMMENTS FROM THE EVENT PERSPECTIVE. REPORT FOR THE LANGUAGE WORK GROUP OF THE FIRST INTERNATIONAL CONFERENCE ON EVENT PERCEPTION.*

Carol Fowler+ and Brad Rakerd++

Signed and spoken utterances have at least two aspects that are of interest to a perceiver. First of all, they have a physical aspect, the significance of which is given in the lawful relations among utterances, the information-bearing media structured by them, and the perceptual systems of observers and listeners. Secondly, they have a linguistic aspect, the significance of which is given in the conventional or ruleful relations between forms and meaning.[1] In part because our time was limited, and in part because so little work has been done on the conventional significance of events (as opposed to the intrinsic significance [cf. Gibson, 1966][2]), our work group chose to focus on the physical aspect. Nevertheless, it will be seen that we did have a speculative word or two to say about the origins of some linguistic conventions, and we would draw attention to the report of the Event/Cognition group, as well as to Verbrugge's remarks (discussant for the address by Studdert-Kennedy), for more elaborate treatments of this important topic.

Roughly, our daily discussions centered around five topic areas: (1) useful descriptions of signed and spoken events; (2) natural constraints on linguistic form; (3) the origins of some linguistic conventions; (4) the ecology of conversation; and (5) conducting language research from an event perspective. Our review of these topics will highlight what seemed to us to be the obvious applications of the event approach and also its apparent limitations.

USEFUL DESCRIPTIONS OF SIGNED AND SPOKEN EVENTS

We considered the minimal linguistic event to be an utterance, and identified as such anything that a talker (signer) might choose to say (sign). Obviously, this definition is unsatisfactory on a number of grounds; however, it does identify the minimal event of interest as being articulatory (gestural) in origin, and rejects as irrelevant those properties of articulation

*The conference was held June 7-12, 1981, in Storrs, Connecticut. The participants in the Language Work Group were Hollis Fitch, Carol Fowler, Nancy Frishberg, Kerry Green, Harlan Lane, Mark Mandall, Brad Rakerd, Robert Remez, Philip Rubin, Judy Shepard-Kegl, Winifred Strange, Michael Studdert-Kennedy, Betty Tuller, and Jerry Zimmerman.

+Also Dartmouth College.

++Also University of Connecticut.

Acknowledgment. This work was supported by NICHD grant HD-01994 to Haskins Laboratories.

(gesture) that are not intended to have linguistic significance. We first attempted to verify that utterances have the "nested" character of other ecological events and that the nestings are perceived; next we considered how to discover the most useful characterization of utterances for the investigators' purposes of studying them as perceived events.

Signing and Speaking as Nested Events

Natural events are nested in the sense that relatively slower, longer-term or more global events are composed of relatively faster, shorter-term or more local ones. For example, a football game is a longer-term event composed of shorter-term plays. It is clear from research--particularly Johansson's (e.g., 1973, 1975) on the perception of form and motion in point-light displays--that viewers are sensitive to the nested structure of events. In his address to this conference, Johansson described an example of light points placed on a rolling wheel. When a single point is affixed to the rim, a viewer who sees only that point gets no sense of the wheel's motion; instead, the percept is of a light moving in a cycloid pattern. However, when a second light is attached, now to the hub of the wheel, the viewer perceives rolling instead of the cycloid motion. Thus, two appropriately placed lights provide sufficient optical information to specify the distal event of rolling.

In geometric terms, rolling involves two kinds of motion: translatory and rotary. These are temporally nested; a series of rotations occurs as the wheel translates over the ground plane. The translatory component affects the behavior of both light points (since both are attached to the translating wheel), but only the point on the rim is affected by the rotary component as well (since it rotates about the point on the hub). Apparently, perceptual sensitivity to the translation (as specified by the correlated activity of the two lights) forms a sort of "backdrop" for detection of the rotation; in essence, the translational component is "factored out" of the cycloid movement of the rim light, thereby revealing its rotational component.

Now let us consider whether these observations apply to signing and its perception. In American Sign Language (ASL), signs are specified by three properties: the shape of the hand or hands, the place of articulation of the sign within a signing space, and the movement of the hand or hands. Signs can be inflected by modulating the movement. For example, a 'distributional' inflection indicating that all of the individuals under discussion are affected by some act is produced by sweeping the arm through the central body plane. By signing, say, GIVE while making such an arm sweep the signer communicates GIVE TO ALL OF THEM. Likewise, a 'temporal' inflection, one indicating the repeated occurrence of an act, is produced by rotating the wrist about a body-centralized point; with this gesture, GIVE is modified to mean GIVE AGAIN AND AGAIN.

Finally, and most importantly for the current discussion, several inflections can be superimposed. Carrying our previous example a step further, it proves possible to sign the complexly inflected verb GIVE TO ALL OF THEM AGAIN AND AGAIN. This is accomplished by rotating the wrist while the arm sweeps through its arc. Notice that when this is done, the optical information for the 'temporal' inflection undergoes a radical transformation; the wrist no longer rotates about a single point fixed at the center of the body, but rather about a point moving with the sweeping arm. It appears that observers treat the sweeping motion (common to all points of the hand, wrist, and arm) as both specifying one signed event (the 'distributional' inflection), and as providing a moving frame of reference for the interpretation of the nested 'temporal' inflection.

Spoken language, with its syntactic units--phonological segments, morphemes, words, and syntactic phrases--and its metrical units--syllables, feet, phonological phrases (see Selkirk, 1980)--lends itself readily to the characterization "nested." We will take an example of nested articulatory and perceived events from a relatively low-level phenomenon, coarticulation. In fluent speech, the productions of successive phonetic segments overlap such that the articulatory gestures often satisfy requirements for two or more segments at the same time. Typically, for example, unstressed vowels coarticulate with the stressed vowels of adjacent syllables. It is therefore tempting to think of the production of the unstressed vowels as being nested within that of their stressed counterparts, and to think of unstressed vowels as being perceived relative to their stressed-vowel context. This way of thinking is promoted by findings (Fowler, 1981) that under some conditions listeners behave as if they have "factored out" the articulatory/acoustic contributions of the context when judging the quality of unstressed vowels--more or less as Johansson's subjects seem to have factored out common and relative motions in an optical display.

In trisyllabic nonsense words with medial /ə/, the medial vowel coarticulates with both of its flanking stressed vowels such that the F2 of /ə/ in, for instance, /ibəbi/ is higher than it is in /ubəbu/. (Compatibly, F2 is high for /i/ and low for /u/.) When extracted from their contexts, the medial /bə/ syllables do sound quite different, but when presented in context they sound alike--more alike, in fact, than do two acoustically *identical* /bə/ syllables presented in different contexts.

A nested-events account of these data would hold that when the /bə/ syllables are extracted from the context in which they had been produced, the perceiver has no way to detect (factor out) the contribution that the stressed vowels have made to that portion of the acoustic signal in which /ə/ correlates predominate over the correlates of other segments--no more than Johansson's subjects can separate the rotary from the translatory components of movements when they see just the one light on the rim of a wheel. Presentation in the context of flanking vowels, on the other hand, allows the perceiver to factor out components in common with those vowels, and to recognize the quality of what is left. This leads to the perceived identity of the acoustically "different" /bə/ syllables (in the /ibəbi/ and /ubəbu/ contexts), and to the perceived difference of the acoustically "identical" syllables (in the different trisyllable contexts).

Identifying Speech Events: The Problem of Description

Several theories of speech perception--including Gibson's (1966) and one more familiar to speech investigators, the motor theory (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967)--adopt a view consistent with an event perspective: namely, that the perceived categories of speech are articulatory in origin. Gibson's view is distinguished from the other by its working assumption that the perceived articulatory categories are fully reflected (however complexly) in the acoustic signal and hence need not be reconstructed by articulatory simulations. What are the reasons for this major disagreement among theorists who agree on the question of what is perceived? One reason may be that they differ in terms of how they describe the acoustic signal or even the articulatory event.

In speech, articulatory activities and their acoustic correlates are both richly structured, and consequently can be described in a great many different ways. Each of the various descriptions may be most appropriate for *certain* purposes, but none is privileged for *all* purposes, and just one or a few are privileged for the purposes of understanding what a talker is doing and why a listener perceives what he or she perceives. A theorist who is convinced that the acoustic support for perceptual categories is inadequate may be correct; but, alternatively, she or he may have selected a description of articulatory events and their acoustic correlates that fails to reveal the support.

There are many reasons why a particular description might be inappropriate for aiding our understanding of speech perception and production. It could specify excessive detail (as when, in Putnam's [1973] example, information about the positions and velocities of the elementary particles of a peg and pegboard are invoked to explain why a square peg won't fit in a round hole). Or, for any level of detail, it could be inappropriate because it classifies components in ways that fail to capture the talker's organization of them or the listener's perceived organization. Appropriate descriptions of vocal activity during speech, then, must capture the organization imposed by the talker; those of the acoustic signal must capture those acoustic reflections of the articulatory organization that are responsible for the listener's perception of it.

Appropriate descriptions of perceived articulatory categories. In some time frame, a talker might be said to have raised his larynx (thereby decreasing the volume of the oral cavity), abducted the vocal folds, increased their stiffness, closed the lips, and raised the body of the tongue toward the palate. This description lists a set of apparently separate articulatory acts. In fact, however, the first three of them have the joint effect of achieving voicelessness; these and the next, lip closure, are the principal components of /p/ articulation; and all five acts together are essential to the production of the syllable /pi/. Thus, the aggregate of occurrences in this time frame have a coordinated structure of relations something like the following: {[(larynx raising, vocal cord abduction, vocal cord stiffening)(lip closure)][tongue-body gesture]}.

If an investigator settles for the first description--a list of the activities of individual articulators--then, from his perspective, information about the phonetic segments of an utterance is already absent and he cannot expect to find any evidence of it in the acoustic signal. Consequently, when a perceiver recovers segments in speech, the recovery must be considered reconstructive. Before settling for this conclusion, however, the investigator can try standing back a little from his first perspective on the vocal tract activity and looking for organizations among gestures that were not initially apparent. These organizations will only be revealed from a temporal perspective broad enough that coupled changes among the coordinated structures can be observed. Certainly if there are coordinative articulatory relations among gestures and if the relations have acoustic reflections, then the listener is likely to be sensitive to the coordinated structure, rather than to the unstructured list of gestures from which it is built, for by detecting the structure of the relations among these gestures, the listener detects the talker's structure--here the featural and phonetic segmental structure of the utterance--which is what she or he must do if the utterance is to be understood.

In support of this general approach to phonetic perception, there is some evidence that listeners do perceive aggregates of articulatory acts as if those acts were coordinated segmental structures. One example of this involves the perception of voicing. Following the release of a voiceless stop consonant, the fundamental frequency (f_0) of the voice is relatively high and falls (Halle & Stevens, 1971; Hombert, 1978; Ohala, 1979). Following a voiced stop, f_0 is low and rises. Although the reasons for this differential patterning of f_0 are not fully understood (Hombert, 1978; Hombert, Ohala, & Ewan, 1979), it is generally agreed that it results from the timing of certain laryngeal adjustments and from certain aerodynamic conditions that the talker establishes in maintaining voicelessness or voicing during the production of the consonant (cf. Abramson & Lisker, 1965). That is, the talker does not plan to produce a high falling f_0 contour following release of a /p/. Instead, he plans to maintain voicelessness of the consonant and an unintended consequence of that effort is a pitch perturbation following release. Compatibly, listeners do not normally hear this pitch difference as such (that is, they do not notice a higher pitched vowel following /p/ than /b/). Instead, in the context of a preceding stop, a high falling f_0 contour in a vowel may serve as information for voicelessness of a preceding consonant (Haggard, Ambler, & Callow, 1970; Fujimura, 1971), even though, when removed from the consonantal context, the f_0 contours are perceived as pitch changes (Hombert, 1978; Hombert et al., 1979).

Also suggestive of the perceptual extraction of coordinated articulatory structures are occasions when the perceiver seems to be misled. Ohala (1974, in press) believes that certain historical sound changes can be explained as results of listeners' having failed to recognize some unplanned articulatory consequence as unplanned. An example related to the first one is the development of distinctive tones in certain languages. These languages evolved from earlier versions without tone systems, but with distinctions in voicing between pairs of consonants. Over time, the f_0 difference just described between syllables differing in initial stop voicing became exaggerated and the voicing distinction was lost. Ohala's interpretation of the source of the change is that in these languages listeners tended to hear the f_0 differences on the post-consonantal vowels as if pitch had been a controlled articulatory variable, rather than an uncontrolled consequence of adjustments related to voicing. Therefore, when these individuals produced the vowels, they generated controlled (and larger) differences in f_0 of voiced and voiceless stop-initial syllables. Eventually, because the f_0 differences had become highly distinctive, the now redundant voicing distinction was lost and the words that formerly had differed in voicing of the initial consonant now differed in tone. According to Ohala, this process occurred during the separation of Punjabi from Hindi.

Appropriate descriptions of the acoustic signal. Because very little is known about how a talker organizes articulation, descriptions of the acoustic signal useful for purposes of understanding perception cannot be guided strongly by information about articulatory categories. However, we do know enough to recognize that the usual method of partitioning the acoustic signal into segments or into "cues" can be improved on. Such partitioning often obscures the existence of information for the phonetic segmental structure of speech because the structure of measured acoustic segments is not coextensive with the phonetic structure of the utterance. For one thing, phonetic segments as produced have a time course that measured acoustic segments do not reflect. The component articulatory gestures of a phonetic segment gradually increase in relative prominence over the residual gestures for a preceding segment and consequently the acoustic signal gradually comes to reflect the articulatory character of the new segment more strongly than that of the old one. Thus, phonetic segments are not discrete on the time axis, although they can be identified as mutually separate and serially ordered by tracking the waxing and waning of their predominance in the acoustic signal (cf. Fant, 1960).

Acoustic segments, on the other hand, are discrete. (Such segments are stretches of the acoustic signal bounded by abrupt changes in spectral composition.) An individual acoustic segment spans far less than all of the acoustic correlates of a phonetic segment and, in general, it reflects the overlapping production of several phonetic segments (cf. Fant, 1960). Looking at the signal as a series of discrete acoustic segments, then, obscures another way of looking at it: as a reflection of a series of overlapping phonetic segments successively increasing and declining in prominence.

Partitioning acoustic signals into acoustic segments also promotes assigning separate status to different acoustic "cues" for a phonetic feature, even though such an assignment tends to violate the articulatory fact that many of these cues, no matter how distinct their acoustic properties may be, are inseparable acoustic products of the gestures for a single phonetic segment (Lisker & Abramson, 1964; Abramson & Lisker, 1965). The findings of "trading relations" among acoustically distinctive parts of the speech signal indicate that these cues are not separable for perceivers any more than they can be for talkers. For example, certain pairs of syllables differing on two distinct acoustic dimensions--the duration of a silent interval following frication noise and the presence or absence of formant transitions into the following vocalic segment--are indistinguishable by listeners (Fitch, Halwes, Erickson, & Liberman, 1980). Within limits, a syllable with a long silent interval and no transitions sounds the same as one with a short silent interval and transitions. It is as if the transitions in the second syllable are indistinguishable from the extra silence in the first. A perceptual theory in which this observation is natural and expected is difficult to imagine--unless the theory recognizes that detecting acoustic segments per se is not all there is to perceiving speech. We would argue that the cues in these stimuli are indistinguishable to the degree that they provide information about the same articulatory event. Thus, 24 msec of silence "trades" with the formant transitions because both cues specify production of /p/. It is our view that source-free descriptions of acoustics will never succeed in capturing what a speech event sounds like to a perceiver, because it is information carried in the signal, not the signal itself, that sounds like something.

NATURAL CONSTRAINTS ON LANGUAGE FORM

Shifting perspectives from ongoing articulation and its reflections in proximal stimulation, we considered how, over the long term, properties of the articulators in speech or of the limbs in sign may have shaped linguistic forms. Similarly, we considered how perceptual systems and acoustic or optical media, with their differential tendencies to be structured by various properties of distal events, may have shaped the forms of sign and speech.

Sign has several regular properties suggestive of natural constraints on manual-language forms. One (Battison, 1974; cited in Siple, 1978, and Klima & Bellugi, 1979) takes the form of a symmetry constraint on two-handed signs: if both hands move in the production of a sign, the shapes and movements of the two hands must be the same and symmetrical. This constraint is compatible with anecdotal evidence (from novice piano players, for example), and more recently with experimental evidence (Kelso, Southard, & Goodman, 1979; Kelso, Holt, Rubin, & Kugler, in press) that it is difficult to engage in different activities with the two hands. One reason for this may be a tendency for actors to reduce the number of independently controlled degrees of freedom in complex tasks by organizing structures coordinatively (e.g., Turvey, 1977). Kelso's experiments suggest that the two arms and hands tend to be organized coordinatively even when such an organization would seem unnecessary or even undesirable (Kelso et al., 1979; Kelso et al., in press); when subjects were required to engage in different activities with the two hands or arms, the "different" movements tended to retain similar properties.

A second constraint, called the "Dominance" constraint by Battison, may have a similar origin in general constraints on movement organization. For signs in which just one hand moves and the other hand serves as a base for the movements (a place of articulation), the base hand must either have the same configuration as the moving hand or one of a very limited set of other configurations.

An example of a constraint in spoken languages may be the tendency for syllable structures to respect a "sonority hierarchy" (e.g., Kiparsky, 1979) whereby sonority (roughly, vowel-likeness) increases inward toward the vowel from both syllable edges. Hence, for example, /tr/, a sequence in which sonority increases from left to right, is an acceptable prevocalic sequence, but postvocalically the order must be /rt/.

As for language features owing to properties of perceptual systems and stimulating media, Lindblom's proposed constraints on the evolution of vowel systems provide an example in spoken languages (1980; see also, Bladon & Lindblom, 1981). Lindblom has proposed that vowel systems maximize the perceptual distances among member vowels. Based on estimates of distances among vowels in perceptual space, he succeeds in predicting which vowels will tend to occur across languages in vowel systems of various sizes. This implies a constraint on phonological inventories that perceivers be able to recover distinct phonetic segments when distinct ones are intended. Talkers cannot elect to realize distinct phonetic segments by using articulatory gestures (however distinct they may be themselves) that fail to leave distinguishing traces in the acoustic medium or in the neural medium of perceptual systems. (Analogous articulatory constraints also operate to shape vowel systems. Thus, the relatively densely populated front vowel space and the sparsely populated back vowel space doubtless reflect the relatively greater agility and precision of movement of the tongue tip and blade compared with the tongue body.)

Lane proposed that similar perceptual and articulatory constraints may shape the evolution of sign inventories. Facial expressions provide information in ASL and perceivers tend to focus on a signer's face. This creates a gradient of acuity peaking at the face. According to Siple (1978), signs made well away from the face tend to be less similar one to the other than signs made in its vicinity; in addition, two handed signs made in the periphery are subject to the Symmetry and Dominance constraints just described, which provide redundancy for the viewer who may not see them as clearly as signs produced near the face. Lane suggested that the relative frequency of signs in various locations in signing space might be predicted jointly by the acuity gradient favoring signs located near the face and a work-minimizing constraint favoring signs closer to waist level.

THE ORIGINS OF SOME LINGUISTIC CONVENTIONS

As we noted earlier, the conventional rather than necessary relationship between linguistic forms and their message function is central to the nature of language, freeing linguistic messages from having to refer to the here and now, and thereby allowing past, future, fictional and hypothetical events all to be discussed. For Gibson, this property of language removes it from the class of things that can be directly perceived:3

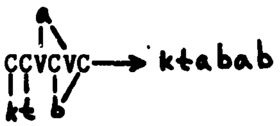
> [Perceptual cognition] is a direct response to things based on stimulus information; [symbolic cognition] is an indirect response to things based on stimulus sources produced by another human individual. The information in the latter is coded; in the former case it cannot properly be called that (1966, p. 91).

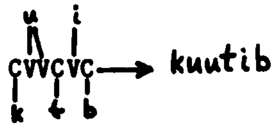
The study group did not discuss language comprehension in relation to event theory, perhaps because event theory currently offers little guidance on that subject. However, there was discussion of the origins of some linguistic conventions. Several examples suggest an origin of certain conventional relations as elaborations of intrinsic ones. The example of tonogenesis given earlier illustrates this idea. Ohala proposes that in some languages distinctive tones originated as controlled exaggerations of the pitch perturbations on vowels caused by the voicing or voicelessness of a preceding consonant.

A second example is so-called "compensatory lengthening" (e.g., Grundt, 1976; Ingria, 1979)--a historical change whereby languages concurrently lost a final consonant in some words and gained a phonological distinction of vowel length, with the words that formerly had ended in a consonant now ending in a phonologically long vowel. In spoken languages, the measured length of vowels shortens when they are spoken before consonants (e.g. Lindblom, Lyberg, & Holmgren, 1981). Of course, since vowels coarticulate with final consonants, this measured shortening may not reflect "true" shortening; presumably, acoustic evidence of their coarticulating edges is obscured by acoustic correlates of the overlaid consonant. In any case, the loss of a final consonant leads to measured lengthening of the vowel. If that unintended lengthening was perceived as controlled lengthening (just as, hypothetically, uncontrolled pitch perturbations were perceived as controlled pitch contours), and was subsequently produced as a controlled lengthening, it could serve as the basis for a phonological distinction in vowel length.

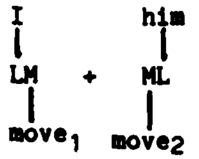
A final example in speech apparently has an analogue in sign. Some speech production investigators have proposed that vowels and consonants are produced by relatively separate articulatory organizations in the vocal tract, and that vowel production may go on essentially continuously during speech production, uninterrupted by concurrently produced consonants (e.g. Öhman, 1966; Perkell, 1969; Fowler, 1980). These proposals are based on observations that vowel-to-vowel gestures that occur during consonant production (Öhman, 1966; Perkell, 1969) sometimes look very similar to vowel-to-vowel gestures in VV sequences (Kent & Moll, 1972). Also, a relatively separate organization of vowel and consonant production with continuous production of vowels may promote such linguistic conventions as vowel infixing in consonantal roots in Arabic languages (McCarthy, 1981) and vowel harmony in languages including Turkish (and in infant babbling [e.g., Menn, 1980]).

Vowel infixing will provide an illustration. In Arabic languages, verb roots are triconsonantal. For example, the root 'ktb' means "write." Verb voice and aspect (e.g., active/passive, perfective/imperfective) are indicated by morphemes consisting entirely of vowels. In McCarthy's recent analysis (1981), the consonantal roots and vowel morphemes are interleaved according to specifications of a limited number of word templates and a small number of principles for assigning the component segments to the templates. Some derivationally related words in Arabic are: <u>katab</u>, <u>ktabab</u>, <u>kutib</u>, and <u>kuutib</u>. The consonantal root in each case is 'ktb'; the vowel morphemes are 'a' (perfective, active) and 'ui' (perfective, passive); and the relevant word templates are CVCVC, CCVCVC, CVVCVC (where C is a consonant and V is a vowel). The general rules for assigning roots and morphemes to templates are (1) to assign the component segments left to right in the template, and (2) if there are more C slots than consonants or more V slots than vowels, to <u>spread</u> the last consonant or last vowel over the remaining C or V slots. The only exception to this generalization is 'i' in 'ui', which is always assigned to the right-most V in the template. Below are two illustrations of verb formation according to this analysis:

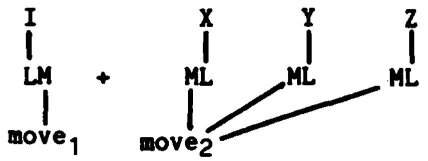




Kegl discussed an analogous system in ASL. A particular root morpheme can be associated with different sign templates to express derivationally or inflectionally related versions of the morpheme. The templates have slots for locations (L) and movements (M), where the former specify person and number and the latter specify aspect. To take an example, the template that underlies I GIVE TO HIM is (LM + ML). Movements and locations are assigned to it as in McCarthy's analysis:



A template can include several L's and M's--more, in fact, than there are distinct movements in a root morpheme. In this case, the movements of the root morpheme are assigned left to right in the template until they are exhausted, and then the right-most movement spreads to fill the empty M slots. In I GIVE TO X, Y, AND Z the template and assignments of root morpheme movements are as follows:



Analyzed this way, the meshing of movements and locations is similar to the meshing of vowels and consonants in languages with infixing and vowel harmony systems. This leads to the question of whether the system is favored as a linguistic device, and, if so, whether it is favored by virtue of the signer's motor organization for producing it. It might be favored, for example, if the motor organization underlying sign production readily produced cyclic repetitions of a movement (as those underlying stepping, breathing, chewing and perhaps vowel production do), and if minimal adjustments to the organization would enable shifts in location without changing the form of the movement.

THE ECOLOGY OF CONVERSATION

A scan of the various conference addresses shows the close ties between the event approach and Gibson's ecological theory of perception. Indeed, Gibson's radical rethinking of classic perceptual problems includes the notion that a perceiver does not operate in a series of "frozen moments," but rather in an ongoing stream of events. We therefore thought it useful to examine the ecology of the speech event, and in doing so we were reminded that both the speaker and the listener (the signer and the observer) have a stake in the success of a communicative episode. This is a rather unique circumstance; it invites both a familiar analysis of the perceiver as an active seeker of information (cf. Gibson, 1966), and a less familiar analysis of the producer as an active provider of informational support.

As to the perceiver's active role, we first of all see behavior intended to enhance signal detection: the head can be rotated to an optimal orientation, the source can be approached, and so on. Beyond this there can be direct communicative intervention; that is, the perceiver can make requests for repetition or clarification. On the producer's part, there are the well-known redundancies of language; in essence, more than enough information is provided to ensure the accuracy of communication. Also, perhaps to avoid syntactic ambiguities, the talker may provide careful prosodic marking for clause boundaries and the like (e.g., Cooper & Paccia-Cooper, 1980). And finally, a talker will enunciate more clearly (and a signer gesture more distinctly) when there is a great distance to the perceiver or when the message context makes a particular word unpredictable.

CONDUCTING LANGUAGE RESEARCH FROM AN EVENT PERSPECTIVE

If there is a theme to the event conference, it is surely that psychologists have paid too little attention to the systematic (and potentially informative) nature of change. With respect to speech, this can be seen in the common practice of decomposing the speech stream into a succession of discrete acoustic segments (e.g., release bursts, aspiration, formant transitions, and the like). A whole literature speaks, in turn, of the difficulty in bringing these acoustic segments into some correspondence with linguistic segments. In the case of sign, the perceptual significance of change was overlooked in early attempts to devise sign glossaries: investigators were preoccupied with cataloguing the featural properties of hand shapes and failed at first to recognize the importance of the gestures being made with the hands (Klima & Bellugi, 1979, chapter 12 and passim; Bellugi & Studdert-Kennedy, 1980).

The members of our group were agreed that a shift of emphasis is needed: investigators of both speech and sign should give greater consideration to the time-varying properties of those events. To begin with, this will involve focusing on the dynamics of the source events themselves. These investigations of the source can suggest compatible and appropriate perceptual analyses. Some recent work using Johansson's point-light techniques to study the coordinated activities of the signer, and the perception of lexical movements and inflections (e.g., Poizner, Bellugi, & Lutes-Driscoll, 1981), seems to offer promising beginnings for such an approach.

Alternatively, analyses of time-varying properties of the signal may provide guidance in understanding the ways in which talkers and signers structure articulatory activity (cf. Fowler, 1979; Tuller & Fowler, 1980). On this issue, our group spent a good deal of time considering the recent work of Remez, Rubin, Pisoni, and Carrell (1981; Remez, Rubin, & Carrell, 1981). They have shown that the phonetic message of an utterance can be preserved in sinewave approximations that reproduce only the center frequencies of its first three formants. These stimuli have no short-time acoustic constituents that vocal tracts can produce and consequently lack many acoustic elements heretofore identified by investigators as speech cues. Presumably the stimuli are intelligible because information is provided by relations among the three sinusoids, information that the sinusoidal variations are compatible with a vocal origin.

These findings are important not because they show short-time acoustic cues to be unimportant to speech perception. After all, naive listeners did not spontaneously hear the sinewaves as phonetic events. Instead, the findings are important in showing that time-varying properties of the signal can provide sufficient information for word and segment identification in speech. In this respect, as Remez and Rubin point out (Note 1), their demonstration is closely analogous to Johansson's demonstrations with point-light displays of moving figures. In both demonstrations, change provides essential information for form.

The conclusion we draw from all of the examples considered here is that students of language should not be misled by the timeless quality of linguistic forms. Signing and speaking are coherent activities and natural classes of events. It is only reasonable to expect that the signatures of these events will be written in time as well as space.

REFERENCE NOTE

1. Remez, R. E., & Rubin, P. E. The stream of speech. Paper distributed at the First International Conference on Event Perception, Storrs, Ct., June, 1981.

REFERENCES

Abramson, A. S., & Lisker, L. Voice onset time in stop consonants: Acoustic analysis and synthesis. In D. E. Commins (Ed.), Proceedings of the 5th International Congress of Acoustics. Liege: Imp. G. Thone, A-51, 1965.

Battison, R. Phonological deletion in American Sign Language. Sign Language Studies, 1974, 5, 1-19.

Bellugi, U., & Studdert-Kennedy, M. Signed and spoken language: Biological constraints on linguistic form. Berlin: Dahlem-Konferenzen, 1980.

Bladon, R., & Lindblom, B. Modeling the judgment of vowel quality differences. Journal of the Acoustical Society of America, 1981, 69, 1414-1422.

Cooper, W. E., & Paccia-Cooper, J. Syntax and speech. Cambridge, Mass.: Harvard University Press, 1980.

Fant, G. Acoustic theory of speech production. Netherlands: Mouton, 1960.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. Perceptual equivalence of two acoustic cues for stop consonant manner. Perception and Psychophysics, 1980, 27, 343-350.

Fowler, C. 'Perceptual centers' in speech production and perception. Perception & Psychophysics, 1979, 25, 375-388.

Fowler, C. Coarticulation and theories of extrinsic timing control. Journal of Phonetics, 1980, 8, 113-133.

Fowler, C. Production and perception of coarticulation among stressed and unstressed vowels. Journal of Speech and Hearing Research, 1981, 46, 127-139.

Fujimura, O. Remarks on stop consonants: Synthesis experiments and acoustic cues. In L. L. Hammerich, R. Jakobson, & E. Zwirner (Eds.), Form and substance: Phonetic and linguistic papers presented to Eli Fischer-Jørgensen. Copenhagen: Akademisk Forlag, 1971.

Gibson, J. J. The senses considered as perceptual systems. Boston, Mass.: Houghton-Mifflin, 1966.

Grundt, A. Compensation in phonology: Open syllable lengthening. Bloomington, Ind.: Indiana University Linguistics Club, 1976.

Haggard, M. P., Ambler, S., & Callow, M. Pitch as a voicing cue. Journal of the Acoustical Society of America, 1970, 47, 613-617.

Halle, M., & Stevens, K. A note on laryngeal features. Quarterly Progress Report, Research Laboratory of Electronics (Massachusetts Institute of Technology), 1971, 101, 198-213.

Hombert, J.-M. Consonant types, vowel quality and tone. In V. Fromkin (Ed.), Tone: A linguistic survey. New York: Academic Press, 1978.

Hombert, J.-M., Ohala, J., & Ewan, W. Phonetic explanation for the development of tones. *Language*, 1979, *55*, 37-58.

Ingria, R. Compensatory lengthening as a metrical phenomenon. *Linguistic Inquiry*, 1979, *11*, 465-495.

Johansson, G. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 1973, *14*, 201-211.

Johansson, G. Visual motion perception. *Scientific American*, 1975, *232*(6), 76-89.

Kelso, J. A. S., Holt, K., Rubin, P., & Kugler, P. Patterns of human interlimb coordination emerge from the properties of nonlinear oscillators: Theory and data. *Journal of Motor Behavior*, in press.

Kelso, J. A. S., Southard, D., & Goodman, D. On the coordination of two-handed movements. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, *5*, 229-238.

Kent, R., & Moll, K. Tongue body articulation during vowel and diphthongal gestures. *Folia Phoniatrica*, 1972, *24*, 278-300.

Kiparsky, P. Metrical structure assignment is cyclic. *Linguistic Inquiry*, 1979, *10*, 421-441.

Klima, E. S., & Bellugi, U. *The signs of language*. Cambridge, Mass.: Harvard University Press, 1979.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431-461.

Lindblom, B. The goal of phonetics and its unification and application. *Phonetica*, 1980, *37*, 7-26.

Lindblom, B., Lyberg, B., & Holmgren, K. *Durational patterns of Swedish phonology: Do they reflect short-term memory processes?* Bloomington, Ind.: Indiana University Linguistics Club, 1981.

Lisker, L., & Abramson, A. S. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 1964, *20*, 384-422.

McCarthy, J. J. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 1981, *12*, 373-418.

Menn, L. Phonological theory and child phonology. In G. Yeni-Komshian, J. F. Kavanaugh, & C. A. Ferguson (Eds.), *Child phonology* (Vol. 1). New York: Academic Press, 1980.

Ohala, J. Experimental historical phonology. In J. M. Anderson & C. Jones (Eds.), *Historical linguistics II: Theory and description in phonology*. Amsterdam: North Holland Publishing Co., 1974.

Ohala, J. The production of tone. In V. Fromkin (Ed.), *Tone: A linguistic survey*. New York: Academic Press, 1979.

Ohala, J. The listener as a source of sound change. In M. F. Miller (Ed.), *Papers from the parasession on language and behavior*. Chicago: Chicago Linguistic Society, in press.

Öhman, S. E. G. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 1966, *39*, 151-168.

Perkell, J. S. *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, Mass.: M.I.T. Press, 1969.

Poizner, H., Bellugi, U., & Lutes-Driscoll, V. Perception of American Sign Language in dynamic point-light displays. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, *7*, 430-440.

Putnam, H. Reductionism and the nature of psychology. *Cognition*, 1973, *2*, 131-146.

Remez, R., Rubin, P., & Carrell, T. Phonetic perception of sinusoidal signals: Effects of amplitude variation. *Journal of the Acoustical*

Society of America, 1981, *69*, S114. (Abstract)

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. Speech perception without traditional speech cues. *Science*, 1981, *212*, 947-950.

Selkirk, E. O. The role of prosodic categories in English word stress. *Linguistic Inquiry*, 1980, *11*, 563-605.

Siple, P. Linguistic and psychological properties of American Sign Language: An overview. In P. Siple (Ed.), *Understanding sign language through sign language research*. New York: Academic Press, 1978.

Tuller, B., & Fowler, C. A. Some articulatory correlates of perceptual isochrony. *Perception & Psychophysics*, 1980, *27*, 277-283.

Turvey, M. T. Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology*. Hillsdale, N.J.: Erlbaum, 1977.

FOOTNOTES

[1] We do not intend to suggest by the word *conventional* that the linguistic aspects of utterances have been established by popular acclaim. We intend only to distinguish the linguistic aspects from the physical aspects in terms of their "relative arbritrariness." Let's consider a physical example first: the articulatory and acoustic differences between the versions of /d/ in /di/ and /du/ are necessary and lawful, given the nature of vocal tracts. This contrasts with the aspiration difference between the versions of /p/ in "pie" and "spy," the production of which is required of English speakers only by convention or rule. We know this to be the case since speakers of other languages (e.g., French) make no such distinction.

[2] In Gibson's view:

> The relation of a perceptual stimulus to its causal source in the environment is of one sort; the relation of a symbol to its referent is of another sort. The former depends on the laws of physics and biology. The latter depends on a *linguistic community*, which is a unique invention of the human species. The relation of perceptual stimuli to their sources is an intrinsic relation such as one of projection, but the relation of symbols to their referents is an extrinsic one of social agreement. The conventions of symbolic speech must be learned, but the child can just about as easily learn one language as another. The connections between stimuli and their sources may well be learned in part, but they make only one language, or better, they do not make a language at all. The language code is cultural, traditional and arbitrary; the connection between stimuli and sources is not (p. 91).

[3] It is interesting in this regard that theories of perception developed within the information-processing framework have relied almost exclusively on verbal materials as stimuli and propose that perception is indirect.

FRICATIVE-STOP COARTICULATION: ACOUSTIC AND PERCEPTUAL EVIDENCE

Bruno H. Repp and Virginia A. Mann+

Abstract. Eight native speakers of American English each produced 10 tokens of all possible CV, FCV, and VFCV utterances with V = [ɑ] or [u], F = [s] or [ʃ], and C = [t] or [k]. Acoustic analysis showed that the formant transition onsets following the stop consonant release were systematically influenced by the preceding fricative, although there were large individual *differences*. In particular, F_3 and F4 tended to be higher following [s] than following [ʃ]. The coarticulatory effects were equally large in FCV (e.g., /stɑ/) and VFCV (e.g., /ɑsdɑ/) utterances; that is, they were not reduced when a syllable boundary intervened between fricative and stop. In a parallel perceptual study, the CV portions of these utterances (with release bursts removed to provoke errors) were presented to listeners for identification of the stop consonant. The pattern of place-of-articulation confusions, too, revealed coarticulatory effects due to the excised fricative context.

INTRODUCTION

In two previous papers (Mann & Repp, 1981; Repp & Mann, 1981) we described an effect of a preceding fricative on stop consonant perception: When a stimulus ambiguous between [tɑ] and [kɑ] was preceded by a fricative noise appropriate for [s] (plus a brief silence appropriate for stop closure), listeners reported [skɑ] more often than [stɑ]. A preceding [ʃ] noise, on the other hand, had little effect on the perceived place of stop articulation. In a series of experiments, we eliminated several possible explanations of the contrasting effects of [s] and [ʃ], such as a simple response bias, auditory contrast, or direct cues to stop place of articulation in the fricative noise. We concluded that the perceptual context effect most likely reflects listeners' expectation of a coarticulatory interaction between a stop consonant and a preceding fricative--namely, a shift in place of stop consonant articulation towards that of the fricative.

+Also Bryn Mawr College.

Acknowledgment. This paper is a revised and expanded version of a paper presented at the 101st Meeting of the Acoustical Society of America in Ottawa, Ontario, May 1981. Our research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to Haskins Laboratories, and by NICHD Postdoctoral Fellowship HD05677 to the second author. We thank Christine Cook and Joyce Schoenheimer who ran subjects and scored data for the perceptual experiment, and Janette Henderson for doing some of the acoustic measurements.

[HASKINS LABORATORIES: Status Report on Speech Research SR-67/68 (1981)]

In our second paper (Repp & Mann, 1981), we reported data that supported this hypothesis. Starting with fricative-stop-vowel utterances obtained from a single speaker, we examined listeners' stop consonant perception after the fricative noise and the stop release burst had been removed. The stops in these truncated CV syllables were more often perceived as having a relatively forward place of articulation when the excised fricative had been [s] than when it had been [ʃ]. In addition, acoustic measurements of the same stimuli showed that the onset frequency of the second formant (F_2) following the stop release was lowered by about 100 Hz in the context of [s], relative to [ʃ] context. A possible difference in F_3 onset in the opposite direction was also indicated. Thus, F_2 and F_3 onsets were more widely separated in [s] context than in [ʃ] context--a pattern that is consistent with the hypothesized forward shift in place of stop articulation following [s], considering the well-known fact that F_2 and F_3 onsets are more widely separated in [tɑ] than in [kɑ].

While these data suggested that fricative-stop coarticulation can occur, their generality was uncertain. In the present paper, we report acoustic measurements and supplementary perceptual tests using utterances collected from eight new speakers.

ACOUSTIC MEASUREMENTS

Method

Speakers. Four males (AA, LL, RM, VG) and four females (VM, SP, PP, FBB), all native speakers of American English, were enlisted. They included two senior phoneticians (AA, LL), an experienced speech scientist (FBB), a graduate student in phonetics (PP), and four speakers with little formal training.

Table 1

The Set of Utterances Used.

[tɑ]	da	[tu]	du
[kɑ]	ga	[ku]	gu
[stɑ]	sta	[stu]	stu
[skɑ]	ska	[sku]	sku
[ʃtɑ]	shta	[ʃtu]	shtu
[ʃkɑ]	shka	[ʃku]	shku
[ɑstɑ]	asda	[ustu]	usdu
[ɑskɑ]	asga	[usku]	usgu
[ɑʃtɑ]	ashda	[uʃtu]	ushdu
[ɑʃkɑ]	ashga	[uʃku]	ushgu

Utterances. The experimental utterances included all possible combinations of an initial vowel ([ɑ], [u], or absent), a fricative ([s], [ʃ], or absent), a stop ([t] or [k]), and a final vowel ([ɑ] or [u]), with the restriction that the two vowels, if present, be the same. Table 1 lists the individual utterances, both in phonetic notation and in the spelling in which they were read by the subjects. Note that the stop consonants, although unaspirated in both FCV and VFCV contexts, were phonologically voiceless in FCV utterances where they were part of a syllable-initial fricative-stop cluster, but phonologically voiced in VFCV utterances where they were in syllable-initial position.[1] Thus, this set of utterances enabled us to assess not only the effect of a preceding fricative on stop articulation but also the sensitivity of that effect to the presence of an intervening syllable boundary.

Ten randomized lists of these utterances were typed on a sheet of paper. The lists included four other utterances ([sɑ], [ʃɑ], [su], and [ʃu]) whose analysis we will not report here. The CV syllables ([tɑ], [kɑ], [tu], [ku]) were added after speakers VM and SP had been recorded; thus, CV data were available for six speakers only.

Recording procedure. The utterances were produced in a soundproof booth in front of a Shure dynamic microphone and recorded on a Crown 800 tape recorder. Speakers were given sample pronunciations by the experimenter and were instructed to read at an even pace and as naturally as possible. Speakers varied in their assignment of stress in the disyllabic (VFCV) utterances: Three (AA, LL, VM) stressed the second syllable while the other five stressed the first syllable. This unintended variation in stress offered the opportunity to observe any possible effects of this variable.

Measurement procedure. Individual utterances were input from audio tape to a Federal UA-6A spectrum analyzer. The results of the spectral analysis were stored in the memory buffer of a GT-40 computer and displayed on a Hewlett-Packard oscilloscope. By using a cursor below a spectrogram of the whole utterance, individual time frames could be selected whose smoothed average spectrum was displayed above the spectrogram, while the corresponding portion of the digitized waveform appeared on a second screen. Thus, the selection of frames for spectral analysis was guided by both waveform and spectrographic information. Spectral cross-sections were computed over a 25.6-msec time frame; the step size from one frame to the next was 12.8 msec. The spectrum was displayed as a point plot with a resolution of 40 Hz. Spectral peaks corresponding to formants were determined from this display by eye and noted down by hand. Appropriate adjustments were made for asymmetric shapes of formant peaks; occasional multiple peaks due to a formant straddling two or more individual harmonics were averaged. In doubtful cases, the spectra of the preceding and following time frames were taken as a guideline.

Because of the laborious nature of this manual procedure, the measurements had to be restricted to the most crucial aspects of the stimuli--the onset frequencies of F_2 and F_3 (and in some cases, F_4) following the stop release. Since the release burst of the stop usually showed a highly irregular spectrum (especially for alveolar stops), it was ignored, and measurements were taken from the first frame that showed a clear formant pattern, normally including F1 (signifying the onset of voicing). Additional

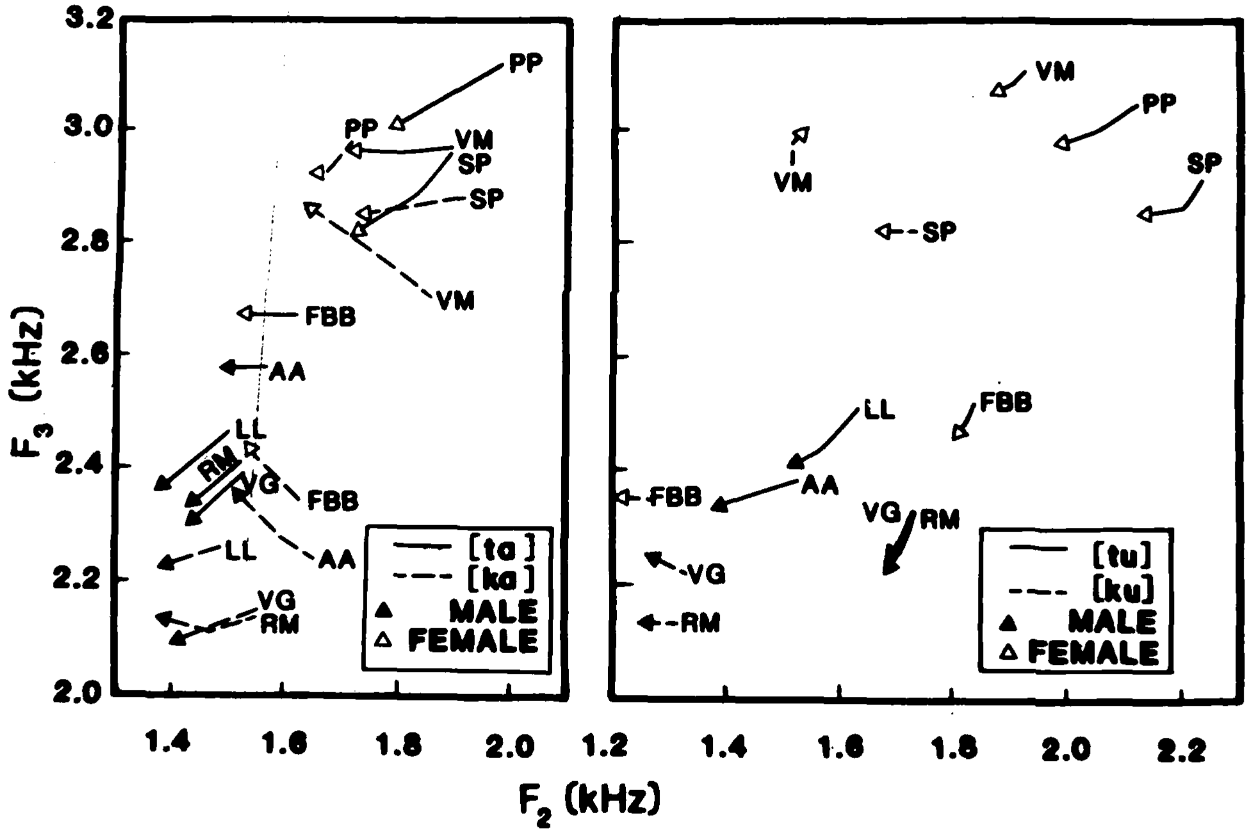


Figure 1. Formant transition patterns for individual speakers' productions of [tɑ], [kɑ], [tu], and [ku], averaged over five different contexts and depicted as trajectories in the F_2-F_3 plane. Data for [ku] are missing from speakers AA, LL, and PP, due to unreliable F_3 measurements.

measurements were taken from the next two frames (only from the next frame in the case of speaker AA whose utterances were the first measured), so that formant transitions were tracked over approximately 50 msec.

Note that this procedure provides a conservative estimate of coarticulatory effects due to the fricative, since any such effects are likely to be most pronounced at the point of stop release and to decrease with distance from the release. Although coarticulatory changes in the release burst may exist (cf. Repp & Mann, 1981, for indirect evidence) they cannot be assessed easily by the present method. Thus, the present investigation was concerned solely with coarticulatory changes in the formant transitions following the release burst.

The raw data consisted of the frequencies of F_2 and F_3 (and, sometimes, F_4) for three (two in the case of AA) consecutive frames of each of ten tokens of 20 utterances (16 in the case of VM and SP) produced by eight speakers. Missing data due to omissions, mispronunciations, or gross acoustic anomalies were rare. A more common source of missing data was the weakness of some formants in certain utterances, particularly F_3 in utterances containing [ku]. For some speakers, as noted below, no reliable data for F_3 could be obtained in these instances.

Results and Discussion

The measurements of F_2 and F_3 in FCV and VFCV utterances were subjected to separate 5-way analyses of variance, with the factors Syllable Boundary (FCV vs. VFCV), Fricative ([s] vs. [ʃ]), Vowel ([ɑ] vs. [u]), Stop ([t] vs. [k]) and Time (3 frames). Speaker AA was not included in these analyses because of *missing data.*

Figure 1 gives an impression of the general frequency characteristics of the formant transitions, regardless of preceding context. The transitions are depicted as trajectories in the F_2-F_3 plane, separately for each speaker's productions of [tɑ], [kɑ], [tu], and [ku], averaged over the five contexts: [-], [s-], [ʃ-], [ɑs-] (or [us-]), and [ɑʃ-] (or [uʃ-]). Except for the few cases with missing data points, each trajectory is based on three points in time separated by 12.8 msec, with 50 measurements per point. In the left panel, it can be seen that all speakers had falling F_2 transitions in both [tɑ] and [kɑ], but two different patterns emerged for F_3: For five speakers (LL, RM, VG, SP, PP), the F_3 transitions were falling for [tɑ] and slightly falling for [kɑ]; for the remaining three speakers (AA, VM, FBB), F_3 was completely flat for [tɑ] but rising for [kɑ]. These individual differences may indicate that the second group of speakers produced [ɑ] with a relatively high F_3. In the right panel, we see that all speakers (except for VM in [ku]) showed falling F_3 transitions in [tu] but a flat F_3 in [ku]. Note that after about 50 msec of formant movement, the formants of [tɑ] and [kɑ], and of [tu] and [ku], were still widely separated, suggesting rather long formant transitions and/or variations in vowel quality dependent on the preceding stop (particularly in [u]).

The trends shown in Figure 1 are all highly significant, and they are generally in agreement with other data in the literature. We will not dwell on them here, as our primary concern was the effect of preceding fricative context. We examined this effect in terms of the difference in formant onset frequencies following [s] and [ʃ].

Table 2 shows these differences (in Hz) for F_2, broken down by individual utterance pairs and speakers but averaged over the three time frames. A positive difference indicates that F_2 was higher following [s] than following [ʃ]. Italics indicate differences that were significant at the $p < .01$ level in individual t-tests. It can be seen that, on the average, F_2 was 4 Hz lower following [s] than following [ʃ]--a nonsignificant difference. Nevertheless, out of 64 individual comparisons, 20 were significant--a proportion far exceeding chance. Of these 20 differences, 8 were positive and 12 negative, which confirms the absence of any general trend. Since there was no pattern in the data, these significant coarticulatory effects must be considered entirely idiosyncratic.

In the analysis of variance, however, there was a significant triple interaction between Fricative, Stop, and Time, F(2,12) = 14.0, $p < .001$: The F_2 transitions of alveolar stops started an average of 40 Hz lower in [s] context than in [ʃ] context, and this difference diminished over time. The F_2 transition of velar stops, on the other hand, was essentially unaffected by fricative context. No other effect involving the Fricative factor was significant, except for one marginally significant 4-way interaction with no clear associated pattern.

The F_3 measurements are shown in Table 3. The picture was quite different here. On the average, F_3 was 46 Hz higher following [s] than following [ʃ], F(1,6) = 51.8, $p < .001$. Of the 61 individual comparisons, 28 were significant, and every single one of them was positive. Thus, even though there was considerable variability across speakers and tokens, the evidence for coarticulatory variation in F_3 is very strong. The correlation between the entries in Tables 2 and 3 is -0.07, indicating no relation between context-induced shifts in F_2 and in F_3.

The coarticulatory effect on F_3 did not decrease over time, suggesting that fricative context may have influenced not only the articulation of the following stop but also that of the following vowel. Two interactions involving the Fricative factor reached significance in the analysis of variance. One--between Fricative, Syllable Boundary, and Time, F(2,12) = 4.2, $p < .05$--revealed that the coarticulatory effect increased over time in FCV utterances but did not change at all over time in VFCV utterances. According to the second interaction--between Fricative, Vowel, Stop, and Time, F(2,12) = 8.0, $p < .01$--the coarticulatory effect increased over time in [u] context and for alveolar stops in [ɑ] context, but decreased over time for velar stops in [ɑ] context. The reasons for these complex patterns are not clear.

Table 4 shows the F_4 measurements, which were obtained for only five speakers and yielded reliable data for only about half the comparisons (mostly those involving stops preceding [u]).[2] Nevertheless, the pattern was very clear: Out of 19 individual comparisons, 18 were positive, and 13 of these were significant. Thus, there was a clear tendency for F_4 to be higher following [s] than following [ʃ]. This tendency seemed to be even stronger than that for F_3, the average difference in Table 4 being more than twice as large (102 Hz) than that in Table 3. However, the changes in F_3 and in F_4 were not significantly correlated (r = 0.21).

Table 2

Coarticulation Effects on F_2: $[F2]_s$- $[F2]_ʃ$ in Hz.

Utterances	Speakers								
	AA	LL	RM	VG	VM	SP	PP	FBB	Mean
[stɑ]-[ʃtɑ]	10	-11	$\underline{-37}$	$\underline{-65}$	32	-24	-21	$\underline{63}$	-7
[skɑ]-[ʃkɑ]	36	-13	1	$\underline{85}$	52	8	0	17	23
[stu]-[ʃtu]	$\underline{98}$	5	$\underline{-64}$	$\underline{73}$	$\underline{-76}$	-12	-47	-44	-8
[sku]-[ʃku]	4	-20	$\underline{76}$	7	49	$\underline{-164}$	-44	$\underline{-147}$	-30
[ɑstɑ]-[ɑʃtɑ]	4	$\underline{-35}$	$\underline{-63}$	$\underline{-57}$	-13	15	-3	-4	-20
[ɑskɑ]-[ɑʃkɑ]	$\underline{131}$	$\underline{51}$	3	-3	$\underline{137}$	44	-33	40	46
[ustu]-[uʃtu]	-22	9	$\underline{-81}$	$\underline{-83}$	-15	4	21	$\underline{-71}$	-30
[usku]-[uʃku]	-10	9	-8	-15	-31	-1	33	-44	-8
Mean	31	-1	-22	-7	17	-16	-12	-24	-4

Note: Underlines indicate difference is significant ($\underline{p}$ < .01) by t-test.

Table 3

Coarticulation Effects on F_3: $[F3]_s$- $[F3]_ʃ$ in Hz.

Utterances	Speakers								
	AA	LL	RM	VG	VM	SP	PP	FBB	Mean
[stɑ]-[ʃtɑ]	-20	$\underline{101}$	(54)	$\underline{43}$	43	37	27	$\underline{117}$	50
[skɑ]-[ʃkɑ]	$\underline{86}$	1	$\underline{76}$	$\underline{61}$	-21	$\underline{64}$	29	$\underline{49}$	43
[stu]-[ʃtu]	$\underline{74}$	$\underline{89}$	$\underline{123}$	$\underline{67}$	28	$\underline{83}$	$\underline{75}$	-9	66
[sku]-[ʃku]		(82)	12	(19)	0	$\underline{71}$	112	11	(44)
[ɑstɑ]-[ɑʃtɑ]	$\underline{54}$	33	-24	$\underline{97}$	12	28	-1	$\underline{145}$	43
[ɑskɑ]-[ɑʃkɑ]	(60)	8	$\underline{104}$	40	-55	11	$\underline{79}$	$\underline{45}$	37
[ustu]-[uʃtu]	$\underline{108}$	$\underline{61}$	15	$\underline{64}$	$\underline{88}$	24	$\underline{125}$	1	61
[usku]-[uʃku]			$\underline{25}$	(9)	-29	25	(46)	$\underline{55}$	(22)
Mean	60	54	48	50	8	43	62	52	46

Note: Underlines indicate difference is significant ($\underline{p}$ < .01) by t-test.
Differences in parentheses are based on a small number of tokens only.

Table 4

Coarticulation Effects on F_4: $[F4]_s$- $[F4]_{ʃ}$ in Hz.

Utterances	Speakers				
	RM	VM	SP	PP	FBB
[stu]-[ʃtu]	35	$\underline{187}$	$\underline{145}$		47
[sku]-[ʃku]		16	$\underline{123}$		
[ɑstɑ]-[ɑʃtɑ]	-1			$\underline{185}$	
[ɑskɑ]-[ɑʃkɑ]	$\underline{79}$			27	
[ustu]-[uʃtu]	$\underline{100}$	36	$\underline{83}$	$\underline{260}$	$\underline{84}$
[usku]-[uʃku]	$\underline{105}$	$\underline{89}$	$\underline{148}$	$\underline{199}$	

Note: Underlines indicate difference is significant ($\underline{p}$ < .01) by t-test.

Table 5

Confusion Matrices for Truncated Stops in [ɑ] and [u] Context.

	Percent Responses									
	V = [ɑ]					V = [u]				
Utterance	"b"	"th"	"d"	"g"	"_"	"b"	"th"	"d"	"g"	"_"
[(s)tV]	16	13	55	10	6	6	5	80	8	1
[(ʃ)tV]	16	9	52	17	6	3	6	80	9	2
[(s)kV]	24	8	21	41	6	63	4	3	19	11
[(ʃ)kV]	26	6	14	46	8	70	3	2	14	11
[(Vs)tV]	6	13	64	9	8	7	3	84	5	1
[(Vʃ)tV]	9	10	63	12	6	3	3	87	6	1
[(Vs)kV]	10	10	32	42	6	52	5	5	29	9
[(Vʃ)kV]	14	8	30	41	7	62	3	4	23	8

A comparison of the F_3 data from each fricative context with the measurements for CV utterances did not confirm our expectation (based on the earlier perceptual data) that the coarticulatory effect would be primarily due to [s]. On the contrary, the data suggest that it was almost entirely due to [ʃ]. However, this difference was in large measure due to a single subject (PP), and because this analysis could be done on five speakers' utterances only, the effects did not reach statistical significance.

We recognize that it is difficult to infer articulatory processes from acoustic data. Given our hypothesis that the place of stop articulation shifts towards that of the preceding fricative (Repp & Mann, 1981), one might expect that the formant transitions of a stop following [s] would be more [t]-like (indicating a forward shift) than those of a stop following [ʃ], which would be more [k]-like (indicating a backward shift). Since [t] has a somewhat higher F_3 onset than [k] in both vocalic contexts (cf. Figure 1), our finding of a higher F_3 onset following [s] is consistent with these expectations. What is not consistent is (1) the absence of any coarticulatory shifts in F_2, particularly in [-u] context where [t] and [k] are characterized by widely differing frequencies (cf. Figure 1), and (2) the finding of higher F_4 onsets following [s], for our data indicate that F_4 is considerably higher in [ku] than in [tu], with less difference between [kɑ] and [tɑ]. In view of these ambiguities, we turned to a perceptual test in the hope that it might shed some light on the direction of the shifts in stop place articulation.

PERCEPTUAL DATA

To complement our acoustic measurements, we gathered perceptual data for a subset of the utterances described above, supposing that labeling responses to FCV and VFCV utterances from which the fricative noise and release burst had been removed might provide another means of assessing any coarticulation between fricative and stop—a procedure used successfully by Repp and Mann (1981). We began by focusing only on those utterances that contained the vowel [ɑ], but later extended our experiment to utterances containing [u].

Method

Subjects. The subjects were ten students from Bryn Mawr and Haverford Colleges, all native speakers of English, of whom eight were paid volunteers and two were participating as part of a class project.

Stimuli. To create the truncated CV syllables, the utterances were digitized at 10 kHz using the Haskins Laboratories PCM system. Individual utterances were displayed on a storage oscilloscope, and the beginning of the first clear pitch pulse following the stop release burst was located in the waveform. Only the stimulus portion following that point was retained. The burst duration (from burst onset to the cutoff point) was recorded. This was done for five tokens of each of all eight speakers' utterances containing the vowel [ɑ], and for four speakers' (AA, LL, PP, FBB) utterances containing the vowel [u].[3]

The truncated CV syllables were assembled into sequences and recorded onto audio tape. A separate tape was created for each speaker and for each vowel, each tape containing 5 repetitions of each of the 40 stimuli (5 tokens of each of 8 utterances) in separately randomized blocks. Interstimulus interval was 2.5 sec, with 7.5 sec between blocks.

Procedure. All subjects participated in two different sessions of approximately one hour. The [ɑ] tapes for speakers LL, VM, RM, and SP were played in the first session and those for speakers AA, PP, VG, and FBB were played in the second session, in the order as listed. Six of the subjects returned for a third session in which all of the [u] tapes were played. The stimuli were presented in a quiet room over TDH-39 earphones. Subjects were required to label each stimulus as containing an initial "b," "th" (as in that), "d," "g," or, if necessary, "-" (no consonant).

Results and Discussion

The data obtained with speaker SP's [ɑ] utterances were excluded from analysis because listeners found it difficult to hear any stops and responded fairly randomly. The combined confusion matrix for the remaining seven speakers' [ɑ] utterances is shown in the left half of Table 5. Comparing utterances differing only in the nature of the original fricative, it is evident that "d" (and "th") responses were somewhat more frequent when the fricative context had been [s], and that "g" (and "b") responses were more frequent when the fricative context had been [ʃ]. Except for the trend in the "b" responses, this pattern is consistent with our hypothesis that [s] leads to a forward shift in the place of articulation of a following stop.

Responses of "d" and "g" were subjected to separate 4-way analyses of variance with the factors Speaker, Stop ([t] vs. [k]), Fricative ([s] vs. [ʃ]), and Syllable Boundary (FCV vs. VFCV). We discovered that, while the effect of fricative context on "d" responses did not reach significance, that on "g" responses did, F(1,9) = 14.5, *p* < .01. However, the extent of this difference varied across speakers, F(6,54) = 8.3, *p* < .001. It was also greater for alveolar stops than for velar ones, F(1,9) = 8.1, *p* < .05, and greater for FCV utterances than for VFCV utterances, F(1,9) = 13.8, *p* < .01. Several other statistical interactions were significant, indicating high variability among utterances produced by different speakers, but consistency in subjects' perception.

To see whether the speaker variability in the perceptual data was related to the similar variability observed in the acoustic measurements, we subtracted the percentage of "g" responses (which had shown a significant effect of fricative context) for each utterance that had contained [s] from that for the corresponding utterance that had contained [ʃ], and then correlated these difference scores (4 values for each of 7 speakers) with the F_3 difference measures of Table 3. The correlation was positive and significant, r(28) = .44, *p* < .02. Thus, pairs of utterances showing a relatively large acoustic effect of fricative context (i.e., higher values of F_3 following [s]) also tended to elicit a larger difference in "g" responses (viz., fewer "g" responses to utterances that originally included [s]).

The confusion matrix for the [u] utterances is shown in the right half of Table 5. There we see that alveolar stops were most often identified as "d," but truncated velar stops received predominantly "b" responses—a finding that may be explained by the similarity of the (equally minimal) formant transitions of labial and velar stops in [u] context (cf. Kewley-Port, 1981), together with a possible listener bias to respond "b" in this context. The table reveals little systematic variation contingent on the excised fricative context, except for a trade between "b" and "g" responses to velar stops: When the preceding fricative had been [s], "b" responses were less frequent, and "g" responses more frequent, than when it had been [ʃ]. These differences, as reflected in the Stop by Fricative interaction, were significant in separate analyses of "b" responses, $F(1,5) = 18.4$, $\underline{p} < .01$, and of "g" responses, $F(1,5) = 15.0$, $\underline{p} < .01$. However, there were a number of significant interactions with other factors, especially with Speakers, reflecting again high between-speaker variability coupled with relatively low between-listener variability. There was no significant correlation with the acoustic measurements for [u] utterances.

CONCLUSIONS

The results of our present studies, even though they are based on a very large amount of data, are not quite as clear as we had hoped. Nevertheless, two conclusions seem appropriate. First, we have obtained rather solid acoustic evidence for a coarticulatory shift in stop production contingent on preceding fricative context. This shift was reflected in generally higher onset valu·s of F_3 and F_4 following [s] than following [ʃ]. Second, we have found additional evidence for fricative-induced shifts in stop production in listeners' perception of the vocalic formant transitions, although the correlation between the acoustic and perceptual findings was weak. Variability of coarticulatory effects across speakers and tokens was unexpectedly large. Unfortunately, neither the acoustic nor the perceptual data have a straightforward articulatory interpretation, which leaves open the question of whether the place of stop articulation indeed *shifts toward that of a preceding* fricative, or whether some more complex articulatory adjustment is involved. Presumably, only direct observations of speech production will shed light on this issue. In our studies, we have laid the foundation for this further research by establishing fricative-stop coarticulation as a real phenomenon in the acoustic and perceptual domains.

REFERENCES

Kewley-Port, D. Representations of spectral change as cues to place of articulation of stop consonants. Unpublished doctoral dissertation, CUNY, 1981.

Mann, V. A., & Repp, B. H. Influence of preceding fricative on stop consonant perception. Journal of the Acoustical Society of America, 1981, 69, 548-558.

Repp, B. H., & Mann, V. A. Perceptual assessment of fricative-stop coarticulation. Journal of the Acoustical Society of America, 1981, 69, 1154-1163.

FOOTNOTES

[1] They were also phonologically voiced in CV utterances, where unspirated [t] and [k] may have alternated with prevoiced [d] and [g]. To simplify the notation, we refer to all stops as [t] or [k].

[2] Average F_4 onset frequencies for five individual speakers (based on a subset of the utterances) were 2862 Hz (RM), 3733 Hz (VM), 3962 Hz (SP), 4303 Hz (PP), and 3626 Hz (FBB).

[3] To check for any possible differences in burst duration contingent on preceding fricative, an analysis of variance was conducted on the burst duration measurements. For the [ɑ] utterances, there was no significant effect of the preceding fricative. Bursts were, however, significantly longer for velar stops (24 msec) than for alveolar ones (16 msec), F(1,7) = 39.2, *p* < .001. Bursts were also significantly longer following a syllable boundary, F(1,7) = 11.3, *p* < .02, although the difference was only 2 msec. In the [u] utterances, too, bursts were longer for velar stops (24 msec) than for alveolar ones (20 msec), F(1,3) = 28.5, *p* < .05, and bursts tended to be longer following [s] (24 msec) than following [ʃ] (20 msec), F(1,3) = 10.7, *p* < .05, both effects being due to unusually short bursts for alveolar stops following [s] (17 msec). The syllable boundary effect was reversed here but nonsignificant.

II. PUBLICATIONS

III. APPENDIX

PUBLICATIONS

Abramson, A. S., Nye, P. W., Henderson, J. B., & Marshall, C. W. Vowel height and the perception of consonantal nasality. *Journal of the Acoustical Society of America*, 1981, *70*, 329-339.

Baer, T. Observation of vocal fold vibration: Measurement of excised larynges. In K. N. Stevens & M. Hirano (Eds.), *Vocal fold physiology*. Tokyo: University of Tokyo Press, 1981, 119-133.

Bell-Berti, F., & Harris, K. S. Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America*, in press.

Bellugi, U., & Studdert-Kennedy, M. (Eds.) *Signed and spoken language: Biological constraints on linguistic form*. Weinheim: Verlag Chemie, 1980.

Fowler, C. A., & Tassinary, L. G. Natural measurement criteria for speech: The anisochrony illusion. In J. Long & A. Baddeley (Eds.), *Attention and performance IX*. Hillsdale, N.J.: Erlbaum, 1981.

Healy, A. F. The effects of visual similarity on proofreading for misspellings. *Memory & Cognition*, 1981, *9*, 453-460.

Henderson, J. B., & Repp, B. H. Is a stop consonant released when followed by another stop consonant? *Phonetica*, in press.

Katz, L., & Baldasare, J. Syllable coding in printed word recognition by children and adults. *Journal of Educational Psychology*, in press.

Liberman, I. Y., & Mann, V. A. Should reading remediation vary with the sex of the child? In A. Ansara, N. Geschwind, A. Galaburda, M. Albert, & N. Gartrell (Eds.), *Sex differences in dyslexia*. Baltimore: The Orton Dyslexia Society, 1981, 151-168.

May, J. G. Acoustic factors that may contribute to categorical perception. *Language and Speech*, 1981, *24*, 273-284.

McGarr, N. S. The effect of context on the intelligibility of hearing and deaf children's speech. *Language and Speech*, 1981, *24*, 255-264.

Metz, D. E., Whitehead, R. L., & McGarr, N. S. Physiological aspects of speech produced by deaf persons. *Audiology: A Journal for Continuing Education*, in press.

Osberger, M. J., & McGarr, N. S. Speech production characteristics of the hearing impaired. In N. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 8). New York: Academic Press, in press.

Remez, R. E., Cutting, J. E., & Studdert-Kennedy, M. Cross-series adaptation using song and string. *Perception & Psychophysics*, 1980, *27*, 524-530.

Remez, R., & Rubin, P. The stream of speech. *Scandinavian Journal of Psychology*, in press.

Repp, B. H. Perceptual equivalence of two kinds of ambiguous speech stimuli. *Bulletin of the Psychonomic Society*, 1981, *18*, 12-14.

Repp, B. H. Two strategies in fricative discrimination. *Perception & Psychophysics*, 1981, *30*, 217-227.

Rubin, P., Baer, T., & Mermelstein, P. An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 1981, *70*, 321-328.

Studdert-Kennedy, M. Cerebral hemispheres: Specialized for the analysis of what? *The Behavioral and Brain Sciences*, 1981, *4*, 76-77.

Studdert-Kennedy, M. The emergence of phonetic structure. *Cognition*, 1981, *10*, 301-306.

Studdert-Kennedy, M. A note on the biology of speech perception. In J. Mehler, M. Garrett, & E. Walker (Eds.), *Perspectives in mental representation*. Hillsdale, N.J.: Erlbaum, in press.

Studdert-Kennedy, M., & Bellugi, U. Introduction. In U. Bellugi & M. Studdert-Kennedy (Eds.), *Signed and spoken language: Biological constraints on linguistic form*. Weinheim: Verlag-Chemie, 1980, 41-56.

Studdert-Kennedy, M., & Lane, H. Clues from the differences between signed and spoken language. In U. Bellugi & M. Studdert-Kennedy (Eds.), *Signed and spoken language: Biological constraints on linguistic form*. Weinheim: Verlag-Chemie, 1980, 29-40.

Verbrugge, R. R. Transformations in knowing: A realist view of metaphor. In R. P. Honeck & R. R. Hoffman (Eds.), *Cognition and figurative language*. Hillsdale, N.J.: Erlbaum, 1980.

Verbrugge, R. R. Two feasts of metaphor. [Review of *Metaphor and thought* by A. Ortony (Ed.), and *On metaphor* by S. Sacks (Ed.).] *Contemporary Psychology*, 1980, *25*, 827-828.

Verbrugge, R. R., & Rakerd, B. Vowel perception: A review of theory and research. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 8). New York: Academic Press, in press.

Verbrugge, R. R., Rakerd, B., Fitch, H., Tuller, B., & Fowler, C. A. The perception of speech events: An ecological perspective. In R. E. Shaw & W. Mace (Eds.), *Event perception*. Hillsdale, N.J.: Erlbaum, in press.

Warren, W. H., & Verbrugge, R. R. Toward an ecological acoustics. In R. E. Shaw & W. Mace (Eds.), *Event perception*. Hillsdale, N.J.: Erlbaum, in press.

Watson, B. C., & Alfonso, P. J. A comparison of LRT and VOT values between stutterers and normal speakers. *Journal of Fluency Disorders*, 1981, in press.

Whalen, D. H. When anaphors are metaphors. In J. Copeland & P. W. Davis (Eds.), *The seventh LACUS forum*. Columbia, SC: Hornbeam Press, 1981, 276-283.

APPENDIX

DTIC (Defense Technical Information Center) and ERIC (Educational Resources Information Center) numbers:

Status Report		DTIC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A013325	ED-109-722
SR-42/43	April - September 1975	AD A018369	ED-117-770
SR-44	October - December 1975	AD A023059	ED-119-273
SR-45/46	January - June 1976	AD A026196	ED-123-678
SR-47	July - September 1976	AD A031789	ED-128-870
SR-48	October - December 1976	AD A036735	ED-135-028
SR-49	January - March 1977	AD A041460	ED-141-864
SR-50	April - June 1977	AD A044820	ED-144-138
SR-51/52	July - December 1977	AD A049215	ED-147-892
SR-53	January - March 1978	AD A055853	ED-155-760
SR-54	April - June 1978	AD A067070	ED-161-096
SR-55/56	July - December 1978	AD A065575	ED-166-757
SR-57	January - March 1979	AD A083179	ED-170-823
SR-58	April - June 1979	AD A077663	ED-178-967
SR-59/60	July - December 1979	AD A082034	ED-181-525
SR-61	January - March 1980	AD A085320	ED-185-636
SR-62	April - June 1980	AD A095062	ED-196-099
SR-63/64	July - December 1980	AD A095860	ED-197-416
SR-65	January - March 1981	AD A099958	ED-201-022
SR-66	April - June 1981	AD A105090	ED-206-038
SR-67/68	July - December 1981	**	**

Information on ordering any of these issues may be found on the following page.

**DTIC and/or ERIC order numbers not yet assigned.

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm International
Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210

Haskins Laboratories Status Report on Speech Research is abstracted in Language and Language Behavior Abstracts, P.O. Box 22206, San Diego, California 92122.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Haskins Laboratories
270 Crown Street
New Haven, Connecticut 06510

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP
N/A

3. REPORT TITLE
Haskins Laboratories Status Report on Speech Research, SR-67/68, July-December, 1981.

4. DESCRIPTIVE NOTES *(Type of report and, inclusive dates)*
Interim Scientific Report

5. AUTHOR(S) *(First name, middle initial, last name)*
Staff of Haskins Laboratories, Alvin M. Liberman, P.I.

6. REPORT DATE
December, 1981

7a. TOTAL NO. OF PAGES
284

7b. NO. OF REFS
514

8a. CONTRACT OR GRANT NO.
HD-01994
HD-05677
N01-HD-1-2420
RR-05596
MCS79-16177
PRF8006144
BNS-8111470
NS13870
NS13617
G-80-0178

9a. ORIGINATOR'S REPORT NUMBER(S)
SR-67/68 (1981)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
None

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited*

11. SUPPLEMENTARY NOTES
N/A

12. SPONSORING MILITARY ACTIVITY
See No. 8

13. ABSTRACT

This report (~~1 July-31 December~~) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

-Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception;
-Temporal patterns of coarticulation: Lip rounding;
-Temporal constraints on anticipatory coarticulation;
-Is a stop consonant released when followed by another stop consonant?
-Obstruent production by hearing-impaired speakers: Interarticulator timing and acoustics;
-On finding that speech is special;
-Reading, prosody, and orthography;
-Children's memory for recurring linguistic and nonlinguistic material in relation to reading ability;
-Phonetic and auditory trading relations between acoustic cues in speech perception: Preliminary results;
-Production and perception of phonetic contrast during phonetic change;
-Decay of auditory memory in vowel discrimination;
-The emergence of phonetic structure;
-Auditory information for breaking and bouncing events: A case study in ecological acoustics;
-Speech and sign: Some comments from an event perspective. Report for the Language Work Group of the First International Conference on Event Perception;
-Fricative-stop coarticulation: Acoustic and perceptual evidence

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED
Security Classification

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech Perception:						
phonetic, auditory, trading relations						
stop consonants, release						
speech processor						
acoustic cues						
vowel discrimination, auditory memory decay						
phonetic structure						
speech, sign language, events						
coarticulation, fricative-stop, acoustic						
Speech Articulation:						
coarticulation, lip rounding, temporal constraints						
obstruents, deaf speakers						
Reading:						
prosody, orthography						
memory, linguistic, nonlinguistic, reading ability						
Ecological Acoustics:						
auditory information, breaking, bouncing, events						

UNCLASSIFIED

Security Classification